Electronic Journal of Statistics
Vol. 9 (2015) 1–79
ISSN: 1935-7524
DOI: 10.1214/15-EJS989

Brittleness of Bayesian inference under finite information in a continuous world

Houman Owhadi* **and Clint Scovel**

California Institute of Technology, MC 9-94, 1200 East California Boulevard, Pasadena, CA 91125, United States of America
e-mail: owhadi@caltech.edu; clintscovel@gmail.com

Tim Sullivan

Mathematics Institute, University of Warwick, Coventry, CV4 7AL, United Kingdom
e-mail: tim.sullivan@warwick.ac.uk

Abstract: We derive, in the classical framework of Bayesian sensitivity analysis, optimal lower and upper bounds on posterior values obtained from Bayesian models that exactly capture an arbitrarily large number of finite-dimensional marginals of the data-generating distribution and/or that are as close as desired to the data-generating distribution in the Prokhorov or total variation metrics; these bounds show that such models may still make the largest possible prediction error after conditioning on an arbitrarily large number of sample data measured at finite precision. These results are obtained through the development of a reduction calculus for optimization problems over measures on spaces of measures. We use this calculus to investigate the mechanisms that generate brittleness/robustness and, in particular, we observe that learning and robustness are antagonistic properties. It is now well understood that the numerical resolution of PDEs requires the satisfaction of specific stability conditions. Is there a missing stability condition for using Bayesian inference in a continuous world under finite information?

MSC 2010 subject classifications: Primary 62F15, 62G35; secondary 62A01, 62E20, 62F12, 62G20.
Keywords and phrases: Bayesian inference, misspecification, robustness, uncertainty quantification, optimal uncertainty quantification.

Received May 2013.

1. Introduction

With the advent of high-performance computing, Bayesian methods are increasingly popular tools for the quantification of uncertainty throughout science and industry. Since these methods impact the making of sometimes critical decisions in increasingly complicated contexts, the sensitivity of their posterior conclusions with respect to the underlying models and prior beliefs is becoming a pressing question.

While it is known that Bayesian methods are robust and consistent when the number of possible outcomes is finite, the exploration of Bayesian inference in

*Corresponding author.

a continuous world has revealed both positive [19, 30, 38, 67, 69, 92, 96] and negative results [12, 13, 35, 47, 48, 61, 71]. One contribution of this paper is the development of a calculus for the elucidation of the mechanisms generating robustness or brittleness in Bayesian inference. In particular, this paper

1. shows that the process of Bayesian conditioning on data at fine enough resolution is sensitive (as defined in [94], modulo a small technicality) with respect to the underlying distributions, under the total variation and Prokhorov metrics; and
2. raises the question of a missing stability condition for using Bayesian inference in a continuous world under finite information, somewhat akin to the CFL condition for the stability of a discrete numerical scheme used to approximate a continuous PDE.

Point (1) is the source of negative results similar to those caused by tail properties in statistics [8, 37], and can be seen as an extreme occurrence of the dilation phenomenon from robust Bayesian inference [103].

Let us now illustrate the main question explored in this paper with a simple example of Bayesian reasoning in action:

Problem 1. There is a bag containing 102 coins, one of which always lands on heads, while the other 101 are perfectly fair. One coin is picked uniformly at random from the bag, flipped 10 times, and 10 heads are obtained. What is the probability that this coin is the unfair coin?

The correct probability is given by applying Bayes' theorem:

$$\mathbb{P}[A|B] = \mathbb{P}[B|A]\frac{\mathbb{P}[A]}{\mathbb{P}[B]} = \frac{1}{1+101\times 2^{-10}} \approx 0.91, \tag{1}$$

where A is the event "the coin is the unfair coin" and B is the event "10 heads are observed". If the number of coins is not known exactly and the supposedly fair coins are not exactly fair, then Bayes' theorem can be used to produce a robust Bayesian inference in the following sense: if the fair coins are slightly unbalanced and the probability of getting a tail is 0.51, and an estimate of 100 coins is used and an estimate $\frac{1}{2}$ of the fairness of the fair coins is used, then the resulting estimate $\frac{1}{1+99\times 2^{-10}}$ is still a good approximation of the correct answer.

Does this robustness hold when the underlying probability space is continuous or an approximation thereof? For example, what if the random outcomes are decimal numbers — perhaps given to finite precision — rather than heads or tails?

1.1. The general question

To investigate these questions in a general context let us now consider the situation in which the space $\mathcal{X}$ where observations/samples take their values is no longer $\{\text{Head}, \text{Tail}\}$ but an arbitrary Polish space (with the real line $\mathbb{R}$ as a

prototypical example). Write $\mathcal{M}(\mathcal{X})$ for the set of probability measures on $\mathcal{X}$ and let $\Phi\colon \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ be a function[1] defining a *quantity of interest*. When $\mathcal{X}$ is the real line $\mathbb{R}$, a prototypical example is $\Phi(\mu) := \mu[X \geq a]$, the probability that the random variable X distributed according to μ exceeds the threshold value a; another typical example is $\Phi(\mu) := \mathbb{E}_\mu[X]$, the mean of X.

Problem 2. Let the *data-generating distribution* $\mu^\dagger \in \mathcal{M}(\mathcal{X})$ be an unknown or partially known probability measure on $\mathcal{X}$. The objective is to estimate $\Phi(\mu^\dagger)$ from the observation of n i.i.d. samples from $\mu^\dagger$, which we denote by $d = (d_1, \ldots, d_n) \in \mathcal{X}^n$.

For practical reasons (and to avoid problems associated with conditioning with respect to events of measure zero) we will assume that the data is observed up to resolution/precision $\delta > 0$, i.e. what we actually observe in Problem 2 is the event $d \in B_\delta^n$, where $B_\delta^n := \prod_{i=1}^n B_\delta(x_i)$, $(x_1, \ldots, x_n)$ is a fixed point of $\mathcal{X}^n$, and $B_\delta(x)$ is the open ball of radius δ and center x (defined with respect to a consistent metric on the Polish space $\mathcal{X}$).

Now observe that the Bayesian answer to Problem 2 is to assume that $\mu^\dagger$ is the realization of some random measure μ on $\mathcal{M}(\mathcal{X})$. This is done by choosing a *model class* $\mathcal{A} \subseteq \mathcal{M}(\mathcal{X})$ and a probability measure $\pi \in \mathcal{M}(\mathcal{A})$ which we call *the prior*. This prior determines the randomness with which a representative $\mu \in \mathcal{A}$ is selected, and for each such $\mu \in \mathcal{A}$, the generation of n i.i.d. samples $d \in \mathcal{X}^n$ by randomly sampling from μ^n naturally determines a product measure on $\mathcal{A} \times \mathcal{X}^n$. In analogy to Problem 1, $\mathcal{A}$ plays the role of the bag of coins (measures) and each measure $\mu \in \mathcal{A}$ plays the role of a coin.

Now the prior estimate of the quantity of interest is $\mathbb{E}_{\mu\sim\pi}[\Phi(\mu)]$ and the posterior estimate is defined as the conditional expectation

$$\mathbb{E}_{\mu\sim\pi, d\sim\mu^n}[\Phi(\mu) | d \in B_\delta^n] \tag{2}$$

with respect to this product measure.

One response to the concern that the choice of prior π is somewhat arbitrary is to explore classes of priors. Indeed:

> "Most statisticians would acknowledge that an analysis is not complete unless the sensitivity of the conclusions to the assumptions is investigated. Yet, in practice, such sensitivity analyses are rarely used. This is because sensitivity analyses involve difficult computations that must often be tailored to the specific problem. This is especially true in Bayesian inference where the computations are already quite difficult." [102]

In this paper we will investigate this approach, known as *robust Bayesian inference* [15, 16, 25, 104] or *Bayesian sensitivity analysis*, and examine the robustness of Bayesian inference by computing optimal bounds on prior and posterior values in terms of given sets of priors. To do so, we need some definitions.

[1] All spaces will be topological spaces, the term "function" will mean Borel measurable function and "measure" will mean Borel measure.

Definition 1.1. For a model class $\mathcal{A} \subseteq \mathcal{M}(\mathcal{X})$, a quantity of interest $\Phi\colon \mathcal{A} \to \mathbb{R}$, and a set of priors $\Pi \subseteq \mathcal{M}(\mathcal{A})$, let

$$\mathcal{L}(\Pi) := \inf_{\pi\in\Pi} \mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\big]$$
$$\mathcal{U}(\Pi) := \sup_{\pi\in\Pi} \mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\big]$$

denote the optimal lower and upper bounds on the prior values of the quantity of interest. For $B \subseteq \mathcal{X}^n$ a non-empty open subset of the data space, let $\Pi_B \subseteq \Pi$ be the subset of priors π such that the probability that $d \in B$ is nonzero, i.e. $\mathbb{P}_{\mu\sim\pi, d\sim\mu^n}[d \in B] > 0$, and let

$$\mathcal{L}(\Pi|B) := \inf_{\pi\in\Pi_B} \mathbb{E}_{\mu\sim\pi, d\sim\mu^n}\big[\Phi(\mu)\big|d \in B\big]$$
$$\mathcal{U}(\Pi|B) := \sup_{\pi\in\Pi_B} \mathbb{E}_{\mu\sim\pi, d\sim\mu^n}\big[\Phi(\mu)\big|d \in B\big]$$

denote the optimal lower and upper bounds on the posterior values of the quantity of interest, given that $d \in B$.

1.2. Example of brittleness under finite information

As illustrated in Problem 1, it is already known from classical Bayesian sensitivity analysis that posterior values are robust if the random outcomes live in a finite space (i.e. $\mathcal{X}$ is finite) or if the class of priors Π is finite-dimensional (i.e. if what one does not know can be represented by a finite number of known parameters). One purpose of this paper is to investigate what the very same classical Bayesian sensitivity analysis framework would conclude in the presence of finite information (i.e. if for instance Π is finite codimensional). To understand this question let us consider the following example:

Example 1.2. Our purpose is to estimate the mean $\Phi(\mu^\dagger) := \mathbb{E}_{\mu^\dagger}[X]$ of some random variable X with respect to some unknown distribution $\mu^\dagger$ on the interval $[0, 1]$ based on the observation of n i.i.d. samples $d := (d_1, \ldots, d_n)$, given to finite resolution δ (i.e. we observe $d \in B^n_\delta$, where B^n_δ is the product of n open balls of radius δ), from the unknown distribution $\mu^\dagger$.

The Bayesian answer to that problem is to assume that $\mu^\dagger$ is the realization of some random measure distributed according to some prior π (i.e. $\mu \sim \pi$) and then compute the posterior value of the mean by conditioning on the data, i.e. compute (2) with $\Phi(\mu) := \mathbb{E}_\mu[X]$. Observe that to specify the prior π we need to specify the distribution of all the moments[2] of μ (i.e. the distribution of the infinite-dimensional vector $(\mathbb{E}_\mu[X], \mathbb{E}_\mu[X^2], \mathbb{E}_\mu[X^3], \ldots)$).

It is known, from classical robust Bayesian inference, that the posterior value (2) is robust with respect to finite dimensional perturbations of the particular choice of the prior π. However, rather than specifying a finite-dimensional

[2]In fact, this is a necessary but not a sufficient condition to determine π, since there are cases in which the moment problem is indeterminate. See [3] for a full discussion of such issues.

class of priors Π (i.e. assuming infinite information), it appears epistemologically more reasonable to specify a finite-*codimensional* Π (i.e. assume finite information) and a natural way to do so is to specify the distribution $\mathbb{Q}$ of only a large, but finite, number of moments of μ (i.e. to specify the distribution of $(\mathbb{E}_\mu[X], \mathbb{E}_\mu[X^2], \ldots, \mathbb{E}_\mu[X^k])$, where $k \in \mathbb{N}$ can be arbitrarily large). This defines a class of priors Π on $\mathcal{M}([0,1])$ such that if $\pi \in \Pi$ and $\mu \sim \pi$ then

$$(\mathbb{E}_\mu[X], \mathbb{E}_\mu[X^2], \ldots, \mathbb{E}_\mu[X^k]) \sim \mathbb{Q}.$$

More precisely, writing Ψ as the function mapping each measure μ on $[0,1]$ to its first k moments $\Psi(\mu) := (\mathbb{E}_\mu[X], \mathbb{E}_\mu[X^2], \ldots, \mathbb{E}_\mu[X^k])$ and choosing a measure $\mathbb{Q}$ on $\Psi(\mathcal{M}([0,1])) \subset \mathbb{R}^k$, Π is simply defined as the pullback of the measure $\mathbb{Q}$ under Ψ, i.e. writing $\mathcal{A} := \mathcal{M}([0,1])$,

$$\Pi := \Psi^{-1}\mathbb{Q} = \{\pi \in \mathcal{M}(\mathcal{A}) \mid \Psi\pi = \mathbb{Q}\}.$$

One consequence of one of the main results of this paper, Theorem 4.13, is that no matter how large k is, no matter how large the number of samples n is, for any $\mathbb{Q}$ that has a density with respect to the uniform distribution on the first k moments, if you observe the data at a fine enough resolution, then the minimum and maximum of the posterior value of the mean over the class of priors Π are 0 and 1, i.e. the following proposition holds.

Proposition 1.3. *For all $k \in \mathbb{N}$, if $\mathbb{Q}$ is absolutely continuous with respect to the uniform distribution on $\Psi(\mathcal{M}([0,1]))$, then*

$$\lim_{\delta \downarrow 0} \mathcal{L}(\Pi | B_\delta^n) = 0 \text{ and } \lim_{\delta \downarrow 0} \mathcal{U}(\Pi | B_\delta^n) = 1$$

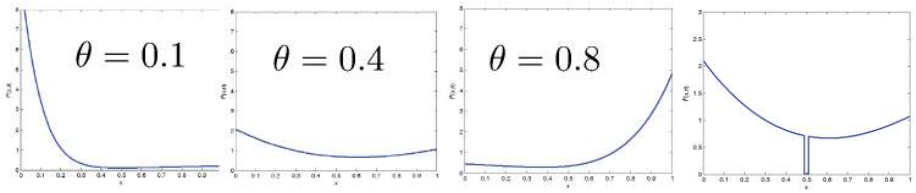
and the convergence holds uniformly in n.

This example of brittleness is derived from Theorem 4.13 (see Example 4.16), the proof of which sheds light on the mechanism leading to brittleness in a general context and shows that the pathology illustrated by Proposition (1.3) is general and inherent to using Bayesian inference in continuous spaces (or their discretizations) under finite information. Furthermore, although this simple example concerns the posterior mean, the quantity of interest in Theorem 4.13 is arbitrary and the brittleness results apply to the whole posterior distribution.

1.3. Example of brittleness under infinitesimal model perturbations

Theorem 4.13 (and its corollary, Theorem 6.1), which leads to brittleness under finite information as illustrated in the previous example, also leads to brittleness under infinitesimal model perturbations in the total variation and Prokhorov metrics. We will now illustrate one mechanism causing brittleness with a simple example.

In this example we are interested in estimating $\Phi(\mu^\dagger) = \mathbb{E}_{\mu^\dagger}[X]$ where $\mu^\dagger$ is an unknown distribution on the unit interval ($\mathcal{X} = [0,1]$) based on the observation



(a) $f^a(x,\theta)$ (b) $f^b(x,\theta)$

FIG 1. *Illustration of the density $f^a(x,\theta)$ of model a and $f^b(x,\theta)$ of model b.*

of a single data point $d_1 = 0.5$ up to resolution δ (i.e. we observe $d_1 \in B_\delta(x_1)$ with $x_1 = 0.5$).

Consider the following two Bayesian models (measures) $\mu^a(\theta)$ and $\mu^b(\theta)$ on the unit interval $[0,1]$, parametrized by $\theta \in (0,1)$. The density f^a of μ^a is given by

$$f^a(x,\theta) = (1-\theta)\big(1+\tfrac{1}{\theta}\big)(1-x)^{\frac{1}{\theta}} + \theta\big(1+\tfrac{1}{1-\theta}\big)x^{\frac{1}{1-\theta}}.$$

The density f^b of μ^b is almost the same as f^a: for $\theta \geq 0.999$, we set $f^b(x,\theta) = f^a(x,\theta)$; but, for $\theta < 0.999$, we set

$$f^b(x,\theta) = f^a(x,\theta)\frac{1}{Z}\big(\mathbb{1}_{\{x\notin(x_1-\frac{\delta_c}{2},x_1+\frac{\delta_c}{2})\}} + 10^{-9}\mathbb{1}_{\{x\in(x_1-\frac{\delta_c}{2},x_1+\frac{\delta_c}{2})\}}\big),$$

where $Z \approx 1$ is a normalization constant so that $\int_0^1 f^b(x,\theta)\,\mathrm{d}x = 1$. See Figure 1 for an illustration of these densities.

Observe that the density of model b is that of model a besides the small gap of width $\delta_c > 0$ created around the data point for model b (if $\theta < 0.999$, see Figure 1); since the data point is fixed at $x_1 = \frac{1}{2}$, the total variation distance $d_{\mathrm{TV}}(\mu^a(\theta),\mu^b(\theta))$ between the two models is, uniformly over $\theta \in (0,1)$, a constant times δ_c. Assuming that the prior distribution on θ is the uniform distribution on $(0,1)$, observe that the prior value of the quantity of interest $\mathbb{E}_\mu[X]$ under both models (a and b) is approximately $\frac{1}{2}$. Now, when θ is close to one (zero) then the density of model a puts most of its mass towards one (zero). Observe also that the density of model b behaves in a similar way, with the important exception that the probability of observing the data under model b is infinitesimally small for $\theta < 0.999$. Therefore, for $\delta < \delta_c$, the posterior value of the quantity of interest $\mathbb{E}_\mu[X]$ under model a is $\frac{1}{2}$ whereas it is close to one under model b. Observe also that a perturbed model c analogous to b would lead to a posterior value close to zero.

This simple example of brittleness under infinitesimal model perturbations is derived from the proof of Theorem 6.4, which shows that Bayesian posterior values are generally brittle under infinitesimal perturbations of Bayesian models in TV and in Prokhorov metrics.

$\mu^b(\theta)$ is also a simple example of what worst priors can look like after a classical Bayesian sensitivity analysis over a class of priors specified via constraints on the TV or Prokhorov distance or the distribution of a finite number of moments.

Can we dismiss these worst priors because they depend on the data? The problem with this argument is that in the context of Bayesian sensitivity analysis worst priors always depend on (or are pre-adapted to) the data. Therefore the same argument would lead to a dismissal of Bayesian sensitivity analysis and therefore the robust Bayesian framework. Can we dismiss these worst priors because they depend *too much* on the data? The problem with this argument is that it is not a transparent task to define *too much* without introducing the following element of circular reasoning: *the degree of pre-adaptation determines the degree of brittleness, the framework is dismissed is when the degree of pre-adaptation is "too much", therefore the method cannot be brittle.*

Can we dismiss these worst priors because they can "look nasty" and make the probability of observing the data very small? The problem with this argument is that these worst priors are not "isolated pathologies" but directions of instability and their number increase with the number of data points. We will illustrate this point with another simple example by placing a uniform constraint on the probability of observing the data in the model class. We already know that if the data is equally likely under all measures in the model class then posterior values are robust but learning is not possible (prior and posterior values are equal). The following example will show that although variations in the probability of the data in the model class make learning possible, they also lead to brittleness.

1.4. Example of learning vs robustness

In this example we are interested in estimating $\Phi(\mu^\dagger) = \mu^\dagger[a,1]$ for some $a \in (0,1)$, where $\mu^\dagger$ is an unknown distribution on the unit interval ($\mathcal{X} = [0,1]$) based on the observation of n data point $d_1, \ldots, d_n$ up to resolution δ (i.e. we observe $d_i \in B_\delta(x_i)$ with $x_i \in [0,1]$ for $i = 1, \ldots, n$).

Our purpose is to examine the sensitivity of the Bayesian answer to this problem with respect to the choice of a particular prior. Consider the model class

$$\mathcal{A} := \mathcal{M}([0,1]), \tag{3}$$

and the class of priors

$$\Pi := \left\{ \pi \in \mathcal{M}(\mathcal{A}) \,\middle|\, \mathbb{E}_{\mu\sim\pi}\left[\mathbb{E}_\mu[X]\right] = m \right\}.$$

Observe that Π corresponds to the assumption that $\mu^\dagger$ is the realization of a random measure on $[0,1]$ whose mean is on average m.

As in the previous example, the finite codimensional class of priors Π leads to brittleness in the sense that the least upper bound on prior values is

$$\mathcal{U}(\Pi) = \frac{m}{a}, \tag{4}$$

whereas, for $\delta \ll 1/n$, the least upper bound on posterior values (using Definition 1.1) is the deterministic supremum of the quantity of interest (over $\mathcal{A}$), i.e.

$$\mathcal{U}(\Pi|B_\delta^n) = 1. \tag{5}$$

Furthermore, worst priors are obtained by selecting priors for which the probability of observing the data $\mu^n[B_\delta^n]$ is arbitrarily close to zero except when $\Phi(\mu)$ is close to its deterministic supremum. The bound on prior values (4) is obtained from theorems 3.6 and 3.11 in Examples 3.7 and 3.15. The bound on posterior values (5) is obtained from theorems 4.8 and 4.13 in Examples 4.9 and 4.16.

Can this brittleness be avoided by adding a uniform constraint on the probability of observing the data in the model class? To investigate this question let us introduce $\alpha \geq 1$ and a probability measure μ_0 on $[0,1]$ with strictly positive Lebesgue density (with a prototypical example being that μ_0 is itself uniform measure on $[0,1]$), consider the (new) model class

$$\mathcal{A}(\alpha) := \left\{ \mu \in \mathcal{M}[0,1] \,\middle|\, \frac{1}{\alpha}\mu_0^n[B_\delta^n] \leq \mu^n[B_\delta^n] \leq \alpha\mu_0^n[B_\delta^n] \right\}, \tag{6}$$

and the (new) class of priors

$$\Pi(\alpha) := \left\{ \pi \in \mathcal{M}(\mathcal{A}(\alpha)) \,\middle|\, \mathbb{E}_{\mu\sim\pi}\left[\mathbb{E}_\mu[X]\right] = m \right\}. \tag{7}$$

Note that, for the model class $\mathcal{A}(\alpha)$, the probability of observing the data is uniformly bounded below by $\frac{1}{\alpha}\mu_0^n[B_\delta^n]$ and above by $\alpha\mu_0^n[B_\delta^n]$. Therefore, for $\alpha = 1$, the probability of observing the data is uniform in the model class, prior values are equal to posterior values, and the method is robust but learning is impossible. If α slightly deviates from 1, then the calculus developed in this paper allows us to compute the least upper bound on posterior values and obtain that

$$\lim_{\delta\to 0} \mathcal{U}\big(\Pi(\alpha)|B_\delta^n\big) = \frac{1}{1+\frac{1}{\alpha^2}\frac{a-m}{m}} = \frac{m}{\frac{a}{\alpha^2}+m(1-\frac{1}{\alpha^2})}. \tag{8}$$

We refer to Example 4.10 for the derivation of (8) from Theorem 4.8.

Note that the right hand side of (8) is equal to m/a for $\alpha = 1$ (when the probability of the data is constant on the model class) and *quickly* converges towards 1 as α increases. As a numerical application observe that for $a = \frac{3}{4}$ and $m = \frac{a}{2} = \frac{3}{8}$, we have $\lim_{\delta\to 0}\mathcal{U}(\Pi(\alpha)) = \frac{1}{2}$ and

$$\lim_{\delta\to 0} \mathcal{U}\big(\Pi(\alpha)|B_\delta^n\big) = \frac{1}{1+\frac{1}{\alpha^2}}$$

Therefore, for $\alpha = 2$, we have (irrespective of the number of data points)

$$\lim_{\delta\to 0} \mathcal{U}\big(\Pi(2)|B_\delta^n\big) = 0.8,$$

and for $\alpha = 10$, we have (irrespective of the number of data points)

$$\lim_{\delta\to 0} \mathcal{U}\big(\Pi(10)|B_\delta^n\big) \approx 0.99.$$

Moreover, if α is derived by assuming the probability of each data point to be known up to some tolerance γ, i.e. if the model class $\mathcal{A}(\alpha)$ is replaced by

$$\mathcal{A}_\gamma := \left\{ \mu \in \mathcal{M}[0,1] \,\middle|\, \frac{1}{\gamma}\mu_0[B_\delta(x_i)] \leq \mu[B_\delta(x_i)] \leq \gamma\mu_0[B_\delta(x_i)] \text{ for } i = 1, \ldots, n \right\} \tag{9}$$

for some $\gamma > 1$, then it can be shown that

$$\lim_{\delta \to 0} \mathcal{U}(\Pi | B_\delta^n) = \frac{1}{1 + \frac{1}{\gamma^{2n}}},$$

which exponentially converges towards 1 as the number n of data points goes to infinity.

In conclusion, the effects of a uniform constraint on the probability of the data under finite information in the model class show that learning ability comes at the price of loss in stability in the following sense: when $\alpha = 1$, the data is equiprobable under all measures in the model class, posterior values are equal to prior values, the method is robust but learning is not possible. As α deviates from one, the learning ability increases as robustness decreases, and when α is large, learning is possible but the method is brittle.

1.5. Missing stability condition for using Bayesian inference under finite information

The previous examples have shown that Bayesian inference can be unstable under finite information, therefore, at the very least, the question of the existence and of the nature of a stability condition for using Bayesian inference remains to be answered. Indeed it is well known that numerical solutions of PDEs can become unstable if specific stability conditions such as the CFL stability condition are not satisfied. Although numerical schemes that do not satisfy the CFL condition may look grossly inadequate, the existence of such perverse examples does not imply the dismissal of the necessity of a stability condition. Similarly, although one may, as in Subsection 1.3, exhibit grossly perverse worst priors, the existence of such priors does not invalidate the question of the missing stability condition for using Bayesian inference under finite information.

The example provided in Subsection 1.4 suggests that, in the framework of Bayesian sensitivity analysis, (i) such a stability condition would depend on how well the probability of the data is known or constrained in the model class, and (ii) learning and robustness are antagonistic/conflicting requirements — there is no free lunch and increased learning potential is paid for by decreased stability of posterior values.

Could this stability condition be derived from closeness in Kullback–Leibler divergence? The problem with this approach is that closeness in Kullback–Leibler divergence cannot be tested with discrete data and it requires the non-singularity of the data generating distribution with respect to the model, which could be a strong assumption for the certification the safety of a critical system.

Indeed, when performing Bayesian analysis on function spaces, as is now increasingly popular, for studying PDE solutions, results like the Feldman–Hájek theorem [45, 56] tell us that *most* pairs of measures are mutually singular, and hence at Kullback–Leibler *distance* infinity from one another. Another problem with using Kullback–Leibler divergence is that a local sensitivity analysis (in the sense of Fréchet derivatives) of posterior values suggests infinite sensitivity as the number of data point goes to infinity [54] (and this result is valid for the broader class of divergences that includes the Hellinger distance).

A close inspection of some of the cases where Bayesian inference has been successful shows the existence of a non-Bayesian feedback loop on the evaluation of its performance [75, 77, 89]. Therefore one natural question is whether the missing stability condition could be derived by exiting the strict framework of Bayesian analysis/inference. According to Efron [43], without genuine prior information

> "Bayesian calculations cannot be uncritically accepted and should be checked by other methods, which usually means frequentistically."

1.6. Calculus for measures over measures

The results of this paper are derived from a calculus allowing us to solve/reduce optimization problems with variables corresponding to measures over measures over arbitrary Polish spaces. The following assertion of Theorem 3.11 is an example of this calculus.

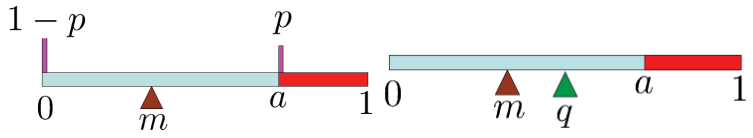
$$\sup_{\pi\in\Psi^{-1}\mathfrak{Q}} \mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\big] = \sup_{\mathbb{Q}\in\mathfrak{Q}} \left[\mathbb{E}_{q\sim\mathbb{Q}}\Big[\sup_{\mu\in\Psi^{-1}(q)} \Phi(\mu)\Big]\right]. \tag{10}$$

In (10), Ψ is a measurable function mapping $\mathcal{A}$ (a Suslin subset of the set $\mathcal{M}(\mathcal{X})$ of probability measures on a Polish space $\mathcal{X}$) into a separable metrizable space $\mathcal{Q}$, $\mathfrak{Q}$ is a subset of $\mathcal{M}(\mathcal{Q})$, and Φ is a measurable quantity of interest defined on $\mathcal{M}(\mathcal{X})$. Therefore, (10) states that the optimization problem (in its left hand side) over $\Psi^{-1}\mathfrak{Q}$ (a subset of the set of measures of $\mathcal{A}$, i.e. a subset of the set of measures of the set of measures over $\mathcal{X}$) is equal to the nesting of an optimization problem over $\Psi^{-1}(q)$ (a subset of $\mathcal{A}$, i.e. a subset of the set of measures over $\mathcal{X}$) and an optimization problem over $\mathfrak{Q}$ (a subset of the set of measures over $\mathcal{Q}$).

We will now illustrate this calculus by showing how (4) can be derived through a simple application of (10). First we need to give a short reminder on optimization over measures via the following problem.

Problem 3. A child is given one pound of playdoh and the seesaw illustrated by Figure 2(a). How much mass can she put above the threshold a while keeping the seesaw balanced at m?

The mathematical formulation of the question articulated in Problem 3 is as follows. What is the least upper bound on $\mathbb{P}[X \geq a]$ if $\mathbb{P}$ is an unknown



(a) Problem 3 (b) Problems 4 and 5

FIG 2. *Reduction of optimization problems over measures and over measures over measures.*

(imperfectly known) probability measure on $[0,1]$ having mean m? The answer to this question is

$$\sup_{\mu \in \mathcal{A}} \mu[a,1] \tag{11}$$

where $\mathcal{A}$ is the set of probability measures on $[0,1]$ having mean m. Although (11) is an infinite dimensional optimization problem over measures, it is easy to see that to achieve the maximum, any mass put above a should be placed exactly at a to create minimum leverage towards the right hand side of the seesaw and any mass put below a should be placed at 0 to create maximum leverage towards the left hand side of the seesaw (as illustrated in Figure 2(a)). This simple argument allows to reduce (11) to a simple one dimensional problem whose solution is $\frac{m}{a}$ and corresponds to Markov's inequality. This simple example of reduction calculus has a generalization to spaces of functions and measures [82] and is based on a form of linear programming in spaces of measures. In particular, the calculus developed in [82] uses results of Winkler [107] — which follow from an extension of Choquet theory (see e.g. [84]) by von Weizsäcker and Winkler [97, Corollary 3] to sets of probability measures with generalized moment constraints — and a result of Kendall [64] characterizing cones, which are lattice cones in their own order.

We will now consider the next level of complexity, illustrated by the following two equivalent problems.

Problem 4. 10,000 children are, each, given one pound of playdoh and a seesaw. On average, how much mass can they put above the threshold a while, on average, keeping the seesaws balanced at m?

Problem 5. A child is given one pound of playdoh and a seesaw. What can you say about how much mass she is putting above the threshold a if all you have is the belief that she is keeping the seesaw balanced at m?

The mathematical formulation of problems 4 and 5 is as follows (for Problem 4, replace 10,000 by N and consider the asymptotic limit $N \to \infty$). What is the least upper bound on $\mathbb{E}_{\mu \sim \pi}[\mu[X \geq a]]$ if π is an unknown (imperfectly known) probability measure on $\mathcal{M}([0,1])$ (the set of probability distributions on $[0,1]$) such that $\mathbb{E}_{\mu \sim \pi}[\mathbb{E}_{\mu}[X]] = m$?

The answer to this question is

$$\sup_{\pi\in\Pi}\mathbb{E}_{\mu\sim\pi}\big[\mu[X\geq a]\big], \tag{12}$$

where Π is the set of measures of probability π on the set of measures of probability on $[0,1]$ such that $\mathbb{E}_{\mu\sim\pi}[\mathbb{E}_{\mu}[X]]=m$.

Although (12) is an optimization over measures over measures, the calculus of (10) introduced in Theorem 3.11 allows us to reduce it to the nesting of two optimization problems over measures as follows.

$$\sup_{\pi\in\Pi}\mathbb{E}_{\mu\sim\pi}\big[\mu[X\geq a]\big]=\sup_{\mathbb{Q}\in\mathcal{M}([0,1]):\,\mathbb{E}_{\mathbb{Q}}[q]=m}\mathbb{E}_{q\sim\mathbb{Q}}\left[\sup_{\mu\in\mathcal{M}([0,1]):\,\mathbb{E}_{\mu}[X]=q}\mu[X\geq a]\right]. \tag{13}$$

Observe that (13) is obtained from (10) by taking $\mathcal{X}=[0,1]$, $\Psi(\mu)=\mathbb{E}_{\mu}[X]$, $\mathcal{Q}=[0,1]$ and $\mathfrak{Q}$ as the set of measures of probability on $\mathcal{Q}$ having mean m. In particular, note that in (13), the inner optimization problem involves taking a supremum over all measures μ on $[0,1]$ having mean q and the outer optimization problem involves taking a supremum over the probability distribution of q, i.e. the set of distributions on $[0,1]$ having mean m. Solving the inner optimization problem as described below Problem 3 leads to:

$$\sup_{\pi\in\Pi}\mathbb{E}_{\mu\sim\pi}\big[\mu[X\geq a]\big]=\sup_{\mathbb{Q}\in\mathcal{M}([0,1]):\,\mathbb{E}_{\mathbb{Q}}[q]=m}\mathbb{E}_{q\sim\mathbb{Q}}\left[\min\left(\frac{q}{a},1\right)\right],$$

and solving the outer optimization step gives the following solution.

$$\sup_{\pi\in\Pi}\mathbb{E}_{\mu\sim\pi}\big[\mu[X\geq a]\big]=\frac{m}{a}.$$

1.7. Structure of the paper and main results

This paper is structured as follows:

Section 2 incorporates Bayesian priors into the Optimal Uncertainty Quantification (OUQ) framework [82]. In the OUQ framework, Uncertainty Quantification (UQ) is formulated as an optimization problem (over an infinite-dimensional set of functions and measures) corresponding to extremizing (i.e. finding worst and best case scenarios) probabilities of failure or other quantities of interest, subject to the constraints imposed by the scenarios compatible with the assumptions and information. In this generalization, priors are probability measures on spaces of measures, and computing optimal bounds on prior values (given a set of priors) requires solving problems in which the optimization variables are measures on spaces of measures (the results of this paper can be extended to measures over spaces of measures and functions but, for the sake of simplicity and clarity, we will limit the presentation to measures over measures).

Section 3 shows how such optimization problems can, under general conditions, be reduced to the nesting of two optimization problems over measures, where then we can apply the reduction theorems of [82].

Section 4 provides similar reduction theorems for the computation of optimal bounds on posterior values given a set of priors and the observation of the data. These reduction theorems lead to the brittleness results of Theorems 4.13, 6.4, and 6.9.

Section 5 reviews questions of Bayesian consistency, inconsistency, model misspecification, and robustness through a motivating analysis and interprets the results of this paper in relation to those questions.

Section 6 presents the brittleness under local misspecification results of Theorems 6.4 and 6.9. That is, given a model, Theorem 6.4 provides optimal bounds on posterior values for priors that are at arbitrarily small distance (in the Prokhorov or total variation metrics) from a given model. Theorems 6.4 and 6.9 show that these optimal bounds on posterior values are the essential supremum and infimum of the quantity of interest irrespective of the size of data and of the size of the metric neighborhood around the model. Finally, Section 8 and Appendix contain the proofs.

2. General set-up

2.1. Notation and conventions

Throughout, for a topological space $\mathcal{Y}$, $\mathcal{B}(\mathcal{Y})$ will denote the Borel σ-algebra of subsets of $\mathcal{Y}$ and $\mathcal{M}(\mathcal{Y})$ will denote the space of Borel probability measures generally endowed with the weak topology and the corresponding Borel σ-algebra unless specified otherwise. For an alternative σ-algebra $\Sigma_{\mathcal{Y}}$ of subsets of $\mathcal{Y}$ the set of probability measures on the σ-algebra $\Sigma_{\mathcal{Y}}$ will be denoted $\mathcal{M}(\Sigma_{\mathcal{Y}})$. For a mapping between topological spaces, the term "measurable" will mean Borel measurable unless specified otherwise. Moreover, suprema over the empty set will have the value $-\infty$ and infima over the empty set the value $+\infty$.

2.2. The general problem and the optimal uncertainty quantification (OUQ) framework

Let $\mathcal{X}$ be Polish and Φ be a measurable function mapping $\mathcal{M}(\mathcal{X})$, the set of measures of probability on $\mathcal{X}$, onto the real line $\mathbb{R}$, known as the *quantity of interest*. Let $\mu^\dagger$ be an unknown or imperfectly known probability measure on $\mathcal{X}$. The general problem guiding our presentation will be that of estimating $\Phi(\mu^\dagger)$.

Let $\mathcal{A}$ be an arbitrary subset of $\mathcal{M}(\mathcal{X})$. If $\mathcal{A}$ represents all that is known about $\mu^\dagger$ (in the sense that $\mu^\dagger \in \mathcal{A}$ and that any $\mu \in \mathcal{A}$ could, a priori, be $\mu^\dagger$ given the available information) then [82] shows that the quantities

$$\mathcal{U}(\mathcal{A}) := \sup_{\mu \in \mathcal{A}} \Phi(\mu) \tag{14}$$

$$\mathcal{L}(\mathcal{A}) := \inf_{\mu \in \mathcal{A}} \Phi(\mu) \tag{15}$$

determine the inequality

$$\mathcal{L}(\mathcal{A}) \leq \Phi(\mu^\dagger) \leq \mathcal{U}(\mathcal{A}), \tag{16}$$

to be optimal given the available information $\mu^\dagger \in \mathcal{A}$ as follows: It is simple to see that the inequality (16) follows from the assumption that $\mu^\dagger \in \mathcal{A}$. Moreover, for any $\varepsilon > 0$ there exists a $\mu \in \mathcal{A}$ such that

$$\mathcal{U}(\mathcal{A}) - \varepsilon < \Phi(\mu) \leq \mathcal{U}(\mathcal{A}).$$

Consequently since all that we know about $\mu^\dagger$ is that $\mu^\dagger \in \mathcal{A}$, it follows that the upper bound $\Phi(\mu^\dagger) \leq \mathcal{U}(\mathcal{A})$ is the best obtainable given that information, and the lower bound is optimal in the same sense.

Although the OUQ optimization problems (14) and (15) are extremely large, we have shown in [82], for the more general situation where $\mathcal{A}$ is a set of functions f and measures μ and Φ a function of (f, μ), that an important subclass enjoys significant and practical finite-dimensional reduction properties. First, by [82, Cor. 4.4], although the optimization variables (f, μ) lie in a product space of functions and probability measures, for OUQ problems governed by linear inequality constraints on generalized moments, the search can be reduced to one over probability measures that are products of finite convex combinations of Dirac masses with explicit upper bounds on the number of Dirac masses. Furthermore, in the special case that all constraints are generalized moments of functions of f, the dependency on the coordinate positions of the Dirac masses is eliminated by observing that the search over admissible functions reduces to a search over functions on an m-fold product of finite discrete spaces, and the search over m-fold products of finite convex combinations of Dirac masses reduces to a search over the products of probability measures on this m-fold product of finite discrete spaces [82, Thm. 4.7]. Finally, by [82, Thm. 4.9], using the lattice structure of the space of functions, the search over these functions can be reduced to a search over a finite set.

For the sake of clarity we will now restrict the presentations of our results to the (simpler) situation where the quantity of interest Φ is (solely) a function of an unknown measure μ. As in [82], the results of this paper can be generalized to situations where Φ is a function of (f, μ).

Example 2.1. A classic example, when $\mathcal{X} = \mathbb{R}$ is $\Phi(\mu) := \mu[X \geq a]$ where a is a safety margin. In the certification context one is interested in showing that $\mu^\dagger[X \geq a] \leq \varepsilon$, where ε is a safety certification threshold (i.e. the maximum acceptable $\mu^\dagger$-probability of the system exceeding the safety margin a). If $\mathcal{U}(\mathcal{A}) \leq \varepsilon$, then the system associated with $\mu^\dagger$ is safe even in the worst case scenario (given the information represented by $\mathcal{A}$). If $\mathcal{L}(\mathcal{A}) > \varepsilon$, then the system associated with $\mu^\dagger$ is unsafe even in the best case scenario (given the information represented by $\mathcal{A}$). If $\mathcal{L}(\mathcal{A}) \leq \varepsilon < \mathcal{U}(\mathcal{A})$, then the safety of the system cannot be decided (although we could declare the system to be unsafe due to lack of information).

2.3. Bayesian priors on the admissible set

In the OUQ setting, an assumption of the form $\mu^\dagger \in \mathcal{A}$ was used to derive the optimal inequality (16). This paper will consider the situation in which

one has priors on the admissible set $\mathcal{A}$ and also information in the form of sample data. One of our goals is to analyse the robustness (or brittleness) of Bayesian inference by obtaining optimal bounds on posterior values given local misspecifications. In that context $\mathcal{A}$ can be viewed as a model class, and $\mu^\dagger$, as the realization of a probability measure (the prior) on $\mathcal{A}$. In order to define priors on the space of admissible scenarios, $\mathcal{A}$ needs to be given the structure of a measurable space; i.e. a suitable σ-algebra $\Sigma_{\mathcal{A}}$ on $\mathcal{A}$ must be provided. From now on, we will assume $\mathcal{A}$ to be a Borel subset of the Polish space $\mathcal{M}(\mathcal{X})$, endowed with the Borel σ-algebra for $\mathcal{A}$. We will also refer to a probability measure $\pi \in \mathcal{M}(\Sigma_{\mathcal{A}})$ as a *prior*.

Remark 2.2. The desire to have the Borel measurable structure of a Polish space might seem to be a spurious level of abstraction, but there are many good reasons for it. The first is that, by Suslin's Theorem [63, Thm. 14.2], all Borel subsets of a Polish space are Suslin, where a *Suslin space* is a continuous Hausdorff image of a Polish space. Indeed, Suslin sets are important in measurable selection theorems (see e.g. [29]) such as those that we use in the proof of Lemma 3.10; furthermore, in addition to Ulam's theorem [6, Thm. 4.3.8] that all probability measures on a Polish space are regular (approximable from within by compact sets), Schwartz' theorem [87] implies that that all probability measures on a Suslin space are regular, and, therefore, [95, Thm. 11.1] implies that the extreme points in the space of probability measures on a Suslin space are the Dirac measures. Consequently, when $\mathcal{M}(\mathcal{X})$ is Polish, any Borel subset $\mathcal{A} \subseteq \mathcal{M}(\mathcal{X})$ is Suslin and so the extreme points of probability measures on $\mathcal{A}$ are the Dirac measures, and some powerful measurable selection theorems are available. Moreover, when the base space is metrizable, then the space of probability measures is Polish in the weak topology if and only if the base space is Polish.

Furthermore, since separability is equivalent to second countability for metric spaces, we have that the Borel structure of a product is the product of Borel structures of Polish spaces. In addition, by [40, Thm. 10.2.2], regular conditional probabilities exist for observables with values in a Polish space. Also, Polish spaces are the spaces of Descriptive Set Theory, see e.g. Kechris [63]. Polish spaces appear to be the appropriate spaces to play topological games such as the Banach–Mazur game [83], the Sierpiński game, the Ulam game, the Banach game, and the Choquet game. Moreover, a theorem of Choquet [63, Thm. 8.18] shows that a separable metric space is completely metrizable (and hence Polish) if and only if the second player has a winning strategy in the strong Choquet game. For a review of topological games, see Telgársky's review [93], and for topological games in hyperspace see that of Zsilinszky [108].

2.4. Data spaces and maps

In practice, the probability measure $\mu^\dagger$ is not observed directly; instead the sample data arrives in the form of (realizations of) observation random variables, the distribution of which is related to $\mu^\dagger$. To simplify the current presentation,

we will assume that this relation is determined by a function of $\mu^\dagger$ — such as the case where the data $X_1, \ldots, X_n$ are determined by n independent realizations X_i of the random variable X determined by the possibly unknown distribution $\mu^\dagger$. Throughout this paper we will use the following notation: $\mathcal{D}$ will denote the observable space (i.e. the space in which the sample data take values); $\mathcal{D}$ will be assumed to be a metrizable Suslin space and D will denote a $\mathcal{D}$-valued random variable producing the observed sample data. To represent the dependence of the observation random variable D on the unknown state $\mu^\dagger \in \mathcal{A}$ we introduce a measurable function

$$\mathbb{D}\colon \mathcal{A} \to \mathcal{M}(\mathcal{D}),$$

where $\mathcal{M}(\mathcal{D})$ is given the Borel structure corresponding to the weak topology, to define this relation. The idea is that $\mathbb{D}(\mu)$ is the probability distribution of the observed sample data $D(\mu)$ if $\mu^\dagger = \mu$, and for this reason it may be called the *data map* or — even more loosely — the *observation operator*. Often, for simplicity, we will write D instead of $D(\mu)$. Note that when the data comes in the form of n i.i.d. realizations of $\mu^\dagger$ we have $\mathcal{D} = \mathcal{X}^n$ and $\mathbb{D}(\mu) = \mu^n$ (where μ^n is the n-fold tensorization of μ).

We proceed with a natural generalization of the Campbell measure and Palm distribution associated with a random measure as described in [62] (see also [33, Ch. 13] for a more current treatment). To that end, observe that since $\mathcal{D}$ is metrizable, it follows from [4, Thm. 15.13], that, for any $B \in \mathcal{B}(\mathcal{D})$, the evaluation $\nu \mapsto \nu(B)$, $\nu \in \mathcal{M}(\mathcal{D})$, is measurable. Consequently, the measurability of $\mathbb{D}$ implies that the mapping

$$\widehat{\mathbb{D}}\colon \mathcal{A} \times \mathcal{B}(\mathcal{D}) \to \mathbb{R}$$

defined by

$$\widehat{\mathbb{D}}(\mu, B) := \mathbb{D}(\mu)[B], \quad \text{for } \mu \in \mathcal{A}, B \in \mathcal{B}(\mathcal{D})$$

is a transition function in the sense that, for fixed $\mu \in \mathcal{A}$, $\widehat{\mathbb{D}}(\mu, \cdot)$ is a probability measure, and, for fixed $B \in \mathcal{B}(\mathcal{D})$, $\widehat{\mathbb{D}}(\cdot, B)$ is Borel measurable. Therefore, by [22, Thm. 10.7.2], any $\pi \in \mathcal{M}(\mathcal{A})$, defines a probability measure

$$\pi \odot \mathbb{D} \in \mathcal{M}\big(\mathcal{B}(\mathcal{A}) \times \mathcal{B}(\mathcal{D})\big)$$

through

$$\pi \odot \mathbb{D}\big[A \times B\big] := \mathbb{E}_{\mu\sim\pi}\big[\mathbb{1}_A(\mu)\mathbb{D}(\mu)[B]\big], \quad \text{for } A \in \mathcal{B}(\mathcal{A}), B \in \mathcal{B}(\mathcal{D}), \tag{17}$$

where $\mathbb{1}_A$ is the indicator function of the set A:

$$\mathbb{1}_A(\mu) := \begin{cases} 1, & \text{if } \mu \in A, \\ 0, & \text{if } \mu \notin A. \end{cases}$$

It is easy to see that π is the $\mathcal{A}$-marginal of $\pi \odot \mathbb{D}$. Moreover, when $\mathcal{X}$ is Polish, [4, Thm. 15.15] implies that $\mathcal{M}(\mathcal{X})$ is Polish, and it follows that $\mathcal{A} \subseteq \mathcal{M}(\mathcal{X})$ is

second countable. Consequently, since $\mathcal{D}$ is Suslin and hence second countable, it follows from [40, Prop. 4.1.7] that

$$\mathcal{B}\big(\mathcal{A}\times\mathcal{D}\big)=\mathcal{B}(\mathcal{A})\times\mathcal{B}(\mathcal{D})$$

and hence $\pi\odot\mathbb{D}$ is a probability measure on $\mathcal{A}\times\mathcal{D}$. That is,

$$\pi\odot\mathbb{D}\in\mathcal{M}(\mathcal{A}\times\mathcal{D}).$$

Let us refer to an element of $\mathcal{M}(\mathcal{A})$ as a *prior* on $\mathcal{A}$. With a prior π on $\mathcal{A}$, the quantity of interest $\Phi(\mu)$ becomes a random variable and we will be interested in estimating its distribution conditioned on the observation $D\in B$, where $B\in\mathcal{B}(\mathcal{D})$.

Example 2.3. In the context of Example 2.1, we are interested in estimating the probability (under the prior π) that the system is unsafe, conditioned on the observations $D\in B$, i.e. the conditional expectation

$$(\pi\odot\mathbb{D})\Big[\mu[X\geq a]>\epsilon\Big|D\in B\Big].$$

If D corresponds to observing independent realizations of X, then the observation space $\mathcal{D}$ is $\mathcal{X}^n$ and the measure $\mathbb{D}(\mu)$ is μ^n.

If D is the random variable that results from observing n independent realizations of $(X+\xi)$ (X is observed with additive Gaussian noise $\xi\sim\mathcal{N}(0,\sigma^2)$), then the measure $\mathbb{D}(\mu)$ is the one associated with the random variable $D=(X^1+\xi^1,\ldots,X^n+\xi^n)$ where the X^i are independent and distributed according to μ and the ξ^i are independent Gaussian random variables of mean zero and variance σ^2.

2.5. Bayes' theorem and conditional expectation

Henceforth $\mathcal{A}$ will be a Suslin space, and suppose now that we have $\pi\odot\mathbb{D}\in\mathcal{M}(\mathcal{A}\times\mathcal{D})$ constructed in the above way. Let $\pi\cdot\mathbb{D}$ denote the corresponding Bayes' sampling distribution defined by the $\mathcal{D}$-marginal of $\pi\odot\mathbb{D}$, and note that, by (17), we have

$$\pi\cdot\mathbb{D}[B]:=\mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big],\quad\text{for } B\in\mathcal{B}(\mathcal{D}).\tag{18}$$

Since both $\mathcal{D}$ and $\mathcal{A}$ are Suslin it follows that the product $\mathcal{A}\times\mathcal{D}$ is Suslin. Consequently, [22, Cor. 10.4.6] asserts that regular conditional probabilities exist for any sub-σ-algebra of $\mathcal{B}(\mathcal{A}\times\mathcal{D})$. In particular, the product theorem of [22, Thm. 10.4.11] asserts that product regular conditional probabilities

$$\big(\pi\odot\mathbb{D}\big)|_d\in\mathcal{M}(\mathcal{A}),\quad\text{for } d\in\mathcal{D}$$

exist and that they are $\pi\cdot\mathbb{D}$-a.e. unique.

When we consider $\pi\in\mathcal{M}(\mathcal{A})$ a prior, then this result can be interpreted as the posteriors of Bayes' theorem. However, because such regular conditional

probabilities are only uniquely defined $\pi \cdot \mathbb{D}$-a.e., when a data sample $d \in \mathcal{D}$ arrives such that $\pi \cdot \mathbb{D}[\{d\}] = 0$, a posterior $(\pi \odot \mathbb{D})|_d$ that could be *any* of the $\pi \cdot \mathbb{D}$-a.e.-equal regular conditional probabilities evaluated at d appears to have dubious utility. Indeed, the fact that the regular conditional probabilities are only uniquely defined $\pi \cdot \mathbb{D}$-a.e. suggests that integrals of posteriors over subsets $B \in \mathcal{B}(\mathcal{D})$ such that $\pi \cdot \mathbb{D}[B] > 0$ are the more natural objects. Moreover, the restriction that B be an open set is natural for practical reasons, since conditioning on D lying in an open subset B rather than on its exact value is what one has to do when the sample data can only be observed after rounding error. Furthermore, we will show in Section 4 that if the data d have been sampled from a probability measure $\pi^\dagger \cdot \mathbb{D}$ for some $\pi^\dagger \in \mathcal{M}(\mathcal{A})$ (commonly called a "true prior" in Bayesian statistics) then with $\pi^\dagger \cdot \mathbb{D}$ probability one (on the realization of d), the $\pi^\dagger \cdot \mathbb{D}$-measure of any open set containing d is strictly positive. In other words, $\pi^\dagger \cdot \mathbb{D}$-almost surely, $\pi^\dagger$ (the "true prior") belongs to the random subset of $\mathcal{M}(\mathcal{A})$ defined as the priors $\pi \in \mathcal{M}(\mathcal{A})$ such that $\pi \cdot \mathbb{D}[B] > 0$ for any open set B containing the data d (this subset is randomized through the realization of the data d).

Finally, throughout, we will find it useful to assume that

Assumption 1. Φ is semibounded

in that it is either bounded above or bounded below. Semiboundedness is sufficient to ensure that the integral of Φ with respect to any probability measure exists, possibly with the value ∞ or $-\infty$, and such integrands are sufficient for the reduction theorems of Winkler [107] that we use.

Remark 2.4. Note that the assumption that Φ is semibounded is mostly for convenience since integrands which are not semibounded, like that defining the first moment, can be considered by restricting the space of measures to those measures that have well-defined first moments.

2.6. Incompletely specified priors

In practical situations, (1) the choice of a particular prior on $\mathcal{A}$ involves a degree of arbitrariness that may be incompatible with the certification of rare/critical events, and (2) the definition of such a prior is a non-trivial task if $\mathcal{A}$ is infinite dimensional. For these reasons it is necessary to consider situations in which the prior π is imperfectly known or specified. More precisely, the (lack of) information (or specification) on π can be represented via the introduction of a space Π where the subset $\Pi \subseteq \mathcal{M}(\mathcal{A})$ consists of the set of admissible priors π.

One of our goals in allowing incompletely specified priors is to assess the robustness of posterior Bayesian estimates with respect to the particular choice of priors. More precisely we will compute optimal bounds on $\mathbb{E}_\pi[\Phi]$ when $\pi \in \Pi$ and show how these bounds are affected by the introduction of sample data by computing optimal bounds on $\mathbb{E}_{\pi \odot \mathbb{D}}[\Phi|B]$, for $B \in \mathcal{B}(\mathcal{D})$.

3. Optimal bounds on the prior value

Recall that for a subset $\mathcal{A}$ and a measurable quantity of interest $\Phi\colon \mathcal{A} \to \mathbb{R}$, that under the assumption $\mu^\dagger \in \mathcal{A}$, we have the optimal upper $\mathcal{U}(\mathcal{A})$ and lower $\mathcal{L}(\mathcal{A})$ bounds on the *value* $\Phi(\mu^\dagger)$ of the quantity of interest, defined in (14) and (15) by

$$\mathcal{U}(\mathcal{A}) := \sup_{\mu\in\mathcal{A}} \Phi(\mu)$$
$$\mathcal{L}(\mathcal{A}) := \inf_{\mu\in\mathcal{A}} \Phi(\mu).$$

When we put a prior π on $\mathcal{A}$, we have to define the *value* $\bar{\Phi}(\pi)$ *of the prior* π corresponding to an extended quantity $\bar{\Phi}\colon \mathcal{M}(\mathcal{A}) \to \mathbb{R}$ of interest corresponding to Φ. Disregarding integrability concerns, for a given Φ, let us call the induced function

$$\bar{\Phi}(\pi) := \mathbb{E}_\pi[\Phi], \quad \pi \in \mathcal{M}(\mathcal{A}), \tag{19}$$

the canonical one associated with Φ and abuse notation by denoting the function $\bar{\Phi}$ as Φ. For such a canonical quantity of interest, we call the value $\mathbb{E}_\pi[\Phi]$ the *prior value*, and note that the values

$$\mathcal{U}(\Pi) := \sup_{\pi\in\Pi} \mathbb{E}_\pi\big[\Phi\big] \tag{20}$$
$$\mathcal{L}(\Pi) := \inf_{\pi\in\Pi} \mathbb{E}_\pi\big[\Phi\big] \tag{21}$$

form a natural generalization of the values $\mathcal{U}(\mathcal{A})$ and $\mathcal{L}(\mathcal{A})$. Moreover, in the same way that $\mathcal{U}(\mathcal{A})$ and $\mathcal{L}(\mathcal{A})$ are optimal upper and lower bounds on $\Phi(\mu^\dagger)$ given the information that $(\mu^\dagger) \in \mathcal{A}$, $\mathcal{U}(\Pi)$ and $\mathcal{L}(\Pi)$ are optimal upper and lower bounds on $\mathbb{E}_\pi[\Phi]$ given the information that $\pi \in \Pi$. Of course, for these expressions to be well defined, integrability concerns should be addressed. Indeed, Assumption 1 implies that $\mathbb{E}_\pi[\Phi]$ is well defined for any bounded measure π, possibly with the value ∞ or $-\infty$, and therefore the quantities in (20) and (21) are well defined.

Remark 3.1. The restriction that the the extended quantity of interest corresponding to Φ be canonical is really no restriction, but is assumed only to simplify the presentation and notation. Indeed, there are many important extended quantities of interest that are not affine as functions of the measure π. However, all the ones that we have thought of can be handled by small modifications of the present framework, and their inclusion here would simply complicate the presentation and notation. Moreover, note that many affine non-canonical extended quantities of interest become canonical through simple transformations. For example, when $\Phi_1(\mu) := \mu[X \geq a]$ is a quantity of interest, and the extended quantity of interest is the probability that the system is unsafe, i.e. $\pi(\{\mu \mid \mu[X \geq a] > \varepsilon\})$ where $\{\mu \mid \mu[X \geq a] > \varepsilon\}$ is the set of unsafe μ, then this extended quantity of interest is not canonical with respect to Φ_1. However, by transformation to $\Phi_2 := \mathbb{1}_{\{r|r>\varepsilon\}} \circ \Phi_1$, the extended quantity of interest becomes canonical and $\mathcal{U}(\Pi)$ and $\mathcal{L}(\Pi)$, defined in terms of Φ_2, are optimal upper and lower bounds on the probability that the system is unsafe given the set of priors Π.

3.1. General information bounds on prior values

Let $\delta\colon \mathcal{A} \to \mathcal{M}(\mathcal{A})$ be the mapping of points to unit Dirac measures, where δ_μ denotes the Dirac mass at μ, and, for $\Pi \subseteq \mathcal{M}(\mathcal{A})$, define

$$\mathcal{A}_\Pi := \delta^{-1}\Pi = \{\mu \in \mathcal{A} \mid \delta_\mu \in \Pi\}. \tag{22}$$

That is, $\mathcal{A}_\Pi$ consists of those scenarios μ that are not only admissible in the sense that they lie in $\mathcal{A}$, but are also admissible as a prior in the sense that δ_μ is an element of Π.

With the convention that $\mathcal{U}(\varnothing) := -\infty$ and $\mathcal{L}(\varnothing) := +\infty$, the following theorem shows the relationships among $\mathcal{U}(\mathcal{A})$ and $\mathcal{U}(\mathcal{A}_\Pi)$ as defined by (14), $\mathcal{L}(\mathcal{A})$ and $\mathcal{L}(\mathcal{A}_\Pi)$ as defined by (15), and $\mathcal{U}(\Pi)$ and $\mathcal{L}(\Pi)$ as defined by (20) and (21).

Theorem 3.2. *It holds true that*

$$\mathcal{U}(\mathcal{A}_\Pi) \le \mathcal{U}(\Pi) \le \mathcal{U}(\mathcal{A})$$

and

$$\mathcal{L}(\mathcal{A}) \le \mathcal{L}(\Pi) \le \mathcal{L}(\mathcal{A}_\Pi).$$

Moreover, if $\mathcal{A}_\Pi$ is non-empty, then

$$\mathcal{L}(\mathcal{A}) \le \mathcal{L}(\Pi) \le \mathcal{L}(\mathcal{A}_\Pi) \le \mathcal{U}(\mathcal{A}_\Pi) \le \mathcal{U}(\Pi) \le \mathcal{U}(\mathcal{A}).$$

3.2. Priors specified by marginals

In many settings, probability measures or sets of probability measures are specified through generalized moments or other properties of marginal distributions. To analyse this case, let $\mathcal{Q}$ be a topological space and consider a measurable map $\Psi\colon \mathcal{A} \to \mathcal{Q}$. Let us abuse notation by also denoting the corresponding pushforward of measures $\Psi\colon \mathcal{M}(\mathcal{A}) \to \mathcal{M}(\mathcal{Q})$ by the same symbol Ψ. For a probability measure $\mathbb{Q} \in \mathcal{M}(\mathcal{Q})$, let

$$\Psi^{-1}\mathbb{Q} := \{\pi \in \mathcal{M}(\mathcal{A}) \mid \Psi\pi = \mathbb{Q}\}$$

be the set of probability measures $\pi \in \mathcal{M}(\mathcal{A})$ that push forward to $\mathbb{Q}$. More generally, for a non-empty set $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$, let

$$\Psi^{-1}\mathfrak{Q} := \{\pi \in \mathcal{M}(\mathcal{A}) \mid \Psi\pi \in \mathfrak{Q}\} \tag{23}$$

be the set of probability measures $\pi \in \mathcal{M}(\mathcal{A})$ such that $\Psi\pi \in \mathfrak{Q}$. Now, let $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ be an admissible set of Ψ-marginals. Then the corresponding admissible set of priors is $\Psi^{-1}\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{A})$ and the corresponding objects to be computed are $\mathcal{U}(\Psi^{-1}\mathfrak{Q})$ and $\mathcal{L}(\Psi^{-1}\mathfrak{Q})$ according to (20) and (21).

We will now demonstrate how to reduce the computation of $\mathcal{U}(\Psi^{-1}\mathfrak{Q})$ and $\mathcal{L}(\Psi^{-1}\mathfrak{Q})$ when $\mathfrak{Q}$ is specified by linear inequalities. Later, in Section 3.2.2, we will develop a more powerful *nested* reduction which will provide the foundation for our reduction methods.

Before we begin, we need to introduce some terminology. Following Winkler [107], let $\mathcal{Y}$ be a topological space and let $\mathcal{M} \subseteq \mathcal{M}(\mathcal{Y})$ be a convex set of measures. Let $\mathrm{ext}(\mathcal{M})$ denote the set of extreme points of $\mathcal{M}$ and let the evaluation field $\Sigma(\mathrm{ext}(\mathcal{M}))$ be the smallest σ-algebra of subsets of $\mathrm{ext}(\mathcal{M})$ such that the evaluation map $\nu \mapsto \nu(B)$ is measurable for all $B \in \mathcal{B}(\mathcal{Y})$. Then a measure $\nu \in \mathcal{M}(\mathcal{Y})$ is said to be a *barycenter* of $\mathcal{M}$ if there exists a probability measure p on $\Sigma(\mathrm{ext}(\mathcal{M}))$ such that the *barycentric formula*

$$\nu(B) = \int_{\mathrm{ext}(\mathcal{M})} \nu'(B)\,\mathrm{d}p(\nu'), \quad B \in \mathcal{B}(\mathcal{Y}) \tag{24}$$

holds. Furthermore, the following notion of a *measure affine function* is central to Winkler's [107] reduction theorems, which we use:

Definition 3.3. An extended real-valued function F on $\mathcal{M} \subseteq \mathcal{M}(\mathcal{Y})$ is said to be *measure affine* if, for all $\nu \in \mathcal{M}$ and all probability measures p on $\Sigma(\mathrm{ext}(\mathcal{M}))$ for which the barycentric formula (24) holds, F is p-integrable and

$$F(\nu) = \int_{\mathrm{ext}(\mathcal{M})} F(\nu')\,\mathrm{d}p(\nu').$$

A major consequence of Assumption 1, that Φ is semibounded, is that $\mathbb{E}_\nu[\Phi]$ exists, with possible values ∞ and $-\infty$, for all finite measures ν. As a consequence, by [107, Prop. 3.1], the extended-real-valued function $\nu \mapsto \mathbb{E}_\nu[\Phi]$ is measure affine.

3.2.1. Primary reduction for prior values

Let us consider the computation of

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) = \sup_{\pi \in \Psi^{-1}\mathfrak{Q}} \mathbb{E}_\pi[\Phi] \tag{25}$$

when $\mathfrak{Q}$ is specified by n generalized moment inequalities determined by measurable functions $g_1, \ldots, g_n$. The situation for the lower bound $\mathcal{L}(\Psi^{-1}\mathfrak{Q})$ is the same. That is, let $I_1, \ldots, I_n$ be n closed intervals, allowing semi-infinite intervals $(-\infty, q_i]$ and $[q_i, \infty)$, and define

$$\mathfrak{Q} = \{\mathbb{Q} \in \mathcal{M}(\mathcal{Q}) \,|\, \mathbb{E}_\mathbb{Q}[g_i] \in I_i \text{ for } i = 1, \ldots, n\},$$

where implicit in the definition is that all n integrals exist. Then, by a change of variables, $\mathbb{E}_{\Psi\pi}[g_i] = \mathbb{E}_\pi[g_i \circ \Psi]$ holds if either integral exists (see e.g. [10, Cor. 19.2]), so we conclude that

$$\begin{aligned}
\Psi^{-1}\mathfrak{Q} &:= \{\pi \in \mathcal{M}(\mathcal{A}) \mid \Psi\pi \in \mathfrak{Q}\} \\
&= \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_{\Psi\pi}[g_i] \in I_i \text{ for } i = 1, , n\} \\
&= \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_\pi[g_i \circ \Psi] \in I_i \text{ for } i = 1, \ldots, n\}.
\end{aligned}$$

Hence, $\Psi^{-1}\mathfrak{Q}$ is defined by the n generalized moment inequalities corresponding to $g_i \circ \Psi \colon \mathcal{A} \to \mathbb{R}$ for $i = 1, \ldots, n$. Consequently, since the function $\pi \mapsto \mathbb{E}_\pi[\Phi]$ is measure affine, it follows from the reduction theorems of [82] that we can reduce the supremum on the right-hand side of (25) to the convex combination of $n+1$ Dirac masses. To state the theorem we have just proven, let

$$\Delta(n) := \left\{ \sum_{i=0}^{n} \alpha_i \delta_{\mu_i} \,\middle|\, \mu_i \in \mathcal{A}, \alpha_i \geq 0, \text{for } i = 0, \ldots, n \right\}. \tag{26}$$

be the set of non-negative combinations of $n+1$ Dirac masses. Let the vector I of intervals have components I_i for $i = 1, \ldots, n$, let

$$\Pi(I) := \Psi^{-1}\mathfrak{Q}$$

be defined as above, and consider the subset

$$\Pi(I, n) := \Pi(I) \cap \Delta(n) \subseteq \Pi(I) \tag{27}$$

of those measures which are the $(n+1)$-fold convex combinations of Dirac masses.

Theorem 3.4. *Let $\mathcal{A}$ be Suslin, let $\mathcal{Q}$ be separable and metrizable, and let $\Psi \colon \mathcal{A} \to \mathcal{Q}$ be measurable. Moreover, for n measurable functions $g_1, \ldots, g_n \colon \mathcal{Q} \to \mathbb{R}$ and n closed intervals $I_1, \ldots, I_n$, let*

$$\mathfrak{Q} := \{\mathbb{Q} \in \mathcal{M}(\mathcal{Q}) \,|\, \mathbb{E}_{\mathbb{Q}}[g_i] \in I_i \text{ for } i = 1, \ldots, n\}$$

define the admissible set of Ψ-marginals. Then,

$$\mathcal{U}\big(\Pi(I)\big) = \mathcal{U}\big(\Pi(I, n)\big)$$

where

$$\mathcal{U}\big(\Pi(I, n)\big) = \begin{cases} \sup \sum_{i=0}^{n} \alpha_i \Phi(\mu_i) \\ \text{among } \mu_i \in \mathcal{A},\ \alpha_i \geq 0,\ \sum_{i=0}^{n} \alpha_i = 1 \\ \text{such that } \sum_{i=0}^{n} \alpha_i g_j\big(\Psi(\mu_i)\big) \in I_j \text{ for } j = 1, \ldots, n. \end{cases} \tag{28}$$

Remark 3.5. The freedom to determine intervals I_i, $i = 1, \ldots, n$, is one way to incorporate uncertainty and maintain a reduction to $n+1$ Dirac masses. In particular, by choosing semi-infinite intervals $I_i := (-\infty, q_i]$ we obtain a reduction to $n+1$ Dirac masses for inequality constraints of the form $\mathbb{E}_{\mathbb{Q}}[g_i] \leq q_i$, and by choosing point intervals $I_i := [q_i, q_i]$ we obtain a reduction to $n+1$ Dirac masses for equality constraints of the form $\mathbb{E}_{\mathbb{Q}}[g_i] = q_i$. Moreover, by choosing the interval to be semi-infinite or point interval depending on the index i we obtain a reduction to $n+1$ Dirac masses for mixed equality and inequality constraints.

Theorem 3.4 can be put into a canonical form in the following way: by considering the modified feature map $\Psi' \colon \mathcal{A} \to \mathbb{R}^n$ with components

$$\Psi'_i := g_i \circ \Psi, \quad \text{for } i = 1, \ldots, n,$$

it follows from the above that

$$\Psi^{-1}\mathfrak{Q} = \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_\pi[\Psi'] \in I\}\,.$$

That is, by changing from the feature map Ψ to Ψ' we end up with a constraint set defined by the first moment of the vector function Ψ'. Therefore, let us remove the $'$ from Ψ', and require $\Psi\colon \mathcal{A} \to \mathbb{R}^n$ to be measurable. The following theorem is the canonical form of Theorem 3.4. It is a corollary of Theorem 3.4 for the constraint $\mathbb{E}_\pi[\Psi] \in Z$ when $Z = I$ is a closed rectangle. However, it is true for arbitrary $Z \subseteq \mathbb{R}^n$.

Theorem 3.6. *Let $\mathcal{A}$ be Suslin, let $\Psi\colon \mathcal{A} \to \mathbb{R}^n$ be measurable, let $Z \subset \mathbb{R}^n$, and let*

$$\mathfrak{Q} := \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^n) \,|\, \mathbb{E}_{Q\sim\mathbb{Q}}[Q] \in Z\} \tag{29}$$

be the set of those measures whose first moment belongs to Z. Then, for

$$\Pi(Z) := \Psi^{-1}\mathfrak{Q} = \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_\pi[\Psi] \in Z\} \tag{30}$$

and $\Pi(Z,n) := \Pi(Z) \cap \Delta(n)$, we have

$$\mathcal{U}\big(\Pi(Z)\big) = \mathcal{U}\big(\Pi(Z,n)\big)$$

where

$$\mathcal{U}\big(\Pi(Z,n)\big) = \begin{cases} \sup \sum_{i=0}^n \alpha_i \Phi(\mu_i) \\ \textit{among } \mu_i \in \mathcal{A}, \alpha_i \geq 0, \sum_{i=0}^n \alpha_i = 1 \\ \textit{such that } \sum_{i=0}^n \alpha_i \Psi(\mu_i) \in Z. \end{cases} \tag{31}$$

Example 3.7. Let $\mathcal{X} := [0,1]$, $\mathcal{Q} = \mathbb{R}$ and consider the admissible set $\mathcal{A} := \mathcal{M}([0,1])$, the quantity of interest $\Phi(\mu) := \mu[X \geq a]$ for some $a \in (0,1)$, and the map $\Psi\colon \mathcal{A} \to \mathbb{R}$ defined by $\Psi(\mu) := \mathbb{E}_\mu[X]$. Take as the set of admissible priors π on $\mathcal{A}$ the collection

$$\Pi := \big\{\pi \in \mathcal{M}(\mathcal{A}) \,\big|\, \mathbb{E}_{\mu\sim\pi}\big[\mathbb{E}_\mu[X]\big] = q\big\}$$

for some fixed $q \in (0,a)$. Then we will show that

$$\mathcal{U}(\Pi) = q/a. \tag{32}$$

To that end, observe that since $\mathbb{E}_{\mu\sim\pi}[\mathbb{E}_\mu[X]] = \mathbb{E}_\pi[\Psi]$, it follows that

$$\Pi = \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_\pi[\Psi] = q\}\,,$$

so that Theorem 3.6 implies that we can reduce the optimization in $\mathcal{U}(\Pi)$ to the supremum over $\mu_1, \mu_2 \in \mathcal{A}$, $\alpha \in [0,1]$ of

$$\alpha\mu_1[X \geq a] + (1-\alpha)\mu_2[X \geq a]$$

subject to the constraint

$$\alpha\mathbb{E}_{\mu_1}[X] + (1-\alpha)\mathbb{E}_{\mu_2}[X] = q.$$

Introducing the slack variables $q_1 := \mathbb{E}_{\mu_1}[X]$, $q_2 := \mathbb{E}_{\mu_2}[X]$ and using [82, Thm. 4.1] to reduce this problem further in μ_1, μ_2, we obtain that $\mathcal{U}(\Pi)$ is equal to the supremum over $\alpha \in [0,1]$ and $q_1, q_2 \in [0,1]$ of

$$\alpha \min\{1, \tfrac{q_1}{a}\} + (1-\alpha)\min\{1, \tfrac{q_2}{a}\}$$

subject to the constraint $\alpha q_1 + (1-\alpha)q_2 = q$. Observing that the supremum is achieved at $q_1, q_2 \leq a$, we conclude that $\mathcal{U}(\Pi) = q/a$, establishing (32). Moreover, note that $\mathcal{U}(\Pi) = \mathcal{U}(\mathcal{A}_\Pi)$ for $\mathcal{A}_\Pi$ defined in (22) instead of the general inequality $\mathcal{U}(\mathcal{A}_\Pi) \leq \mathcal{U}(\Pi)$ of Theorem 3.2.

3.2.2. Nested reduction for prior values

The result of Example 3.7 can also be deduced through a *nested* reduction that we will find generally more useful for two reasons. The first is that, in practice, not only is it highly non-trivial to specify a prior on the space $\mathcal{A}$, since it requires quantifying information on an infinite-dimensional space, but it may also be undesirable to do so. Indeed, if an expert does not have a prior on the full space $\mathcal{A}$ but only on some projection $\Psi(\mathcal{A}) = \mathcal{Q}$, then, rather than arbitrarily picking one particular prior on the space $\mathcal{A}$ compatible with the specified prior on $\Psi(\mathcal{A})$, it might be preferable to work with the set of priors on $\mathcal{A}$ specified through such marginals. Our second and main motivation is that, even when we can do the reduction on the primary space $\mathcal{M}(\mathcal{A})$, the reduced space remains so large that it may not be amenable to computation. However with the nested reduction theorems given below, the reduced space becomes computationally manageable for finite-dimensional $\mathcal{Q}$.

Example 3.8. Consider $\Phi(\mu) := \mu[X \geq a]$, where a is thought of as a safety margin, $\Psi(\mu) = (\mathbb{E}_\mu[X], \mathrm{Var}_\mu[X])$, $\mathcal{Q} = \mathbb{R}^2$, and $\mathfrak{Q} = \{\mathbb{Q}\}$, where $\mathbb{Q}$ corresponds to the uniform distribution on $[-1,1] \times [3,4]$. In that example, the expert has only "the prior" that the mean of X with respect to μ is uniformly distributed on $[-1,1]$ and that the variance of X with respect to μ is independent of its mean and uniformly distributed on $[3,4]$. Observe that in this situation $\mathfrak{Q}$ does not uniquely specify a prior $\pi \in \mathcal{M}(\mathcal{A})$ but an infinite-dimensional set of priors $\Psi^{-1}(\mathfrak{Q}) \subseteq \mathcal{M}(\mathcal{A})$ and a robust approach would require assessing the safety of the system under the whole set $\Psi^{-1}(\mathfrak{Q})$ rather than under a particular element π of that set.

Idea of the nested reduction Roughly, the idea of the nested reduction is as follows. To compute (25), consider the induced function

$$\mathcal{U} \circ \Psi^{-1} \colon \mathcal{Q} \to \mathbb{R}$$

defined by

$$\big(\mathcal{U} \circ \Psi^{-1}\big)(q) := \mathcal{U}\big(\Psi^{-1}(q)\big) = \sup_{\mu \in \Psi^{-1}(q)} \Phi(\mu), \quad \text{for } q \in \mathcal{Q},$$

where we use the notation of (14). From this it is natural to consider

$$\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}], \quad \text{for } \mathbb{Q} \in \mathfrak{Q}.$$

Let $\mathbb{Q} \in \mathfrak{Q}$. Then, for any π such that $\Psi\pi = \mathbb{Q}$, it follows that

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] &= \mathbb{E}_{\Psi\pi}[\mathcal{U} \circ \Psi^{-1}] \\ &= \mathbb{E}_{\pi}[\mathcal{U} \circ \Psi^{-1} \circ \Psi]\end{aligned}$$

Unfortunately, it is not true that $\mathcal{U} \circ \Psi^{-1} \circ \Psi = \Phi$; instead it is $(\mathcal{U} \circ \Psi^{-1} \circ \Psi)(\mu) = \sup_{\mu' : \Psi(\mu') = \Psi(\mu)} \Phi(\mu')$. However, if it were true, then we would obtain

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] &= \mathbb{E}_{\Psi\pi}[\mathcal{U} \circ \Psi^{-1}] \\ &= \mathbb{E}_{\pi}[\mathcal{U} \circ \Psi^{-1} \circ \Psi] \\ &= \mathbb{E}_{\pi}[\Phi]\end{aligned}$$

and conclude that

$$\sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] = \sup_{\pi \in \Psi^{-1}\mathfrak{Q}} \mathbb{E}_{\pi}[\Phi] = \mathcal{U}(\Psi^{-1}\mathfrak{Q}).$$

We will show that, despite the fact that $\mathcal{U} \circ \Psi^{-1} \circ \Psi \neq \Phi$, the conclusion

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) = \sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] \tag{33}$$

is still valid, provided that it is interpreted correctly. Heuristically, the reason for this is that the supremum $\sup_{\pi \in \Psi^{-1}\mathfrak{Q}}$ in $\mathcal{U}(\Psi^{-1}\mathfrak{Q})$ is exploring the maximum value of Φ on level sets of Ψ very much like the supremum in $(\mathcal{U} \circ \Psi^{-1})(q) = \sup_{\Psi^{-1}(q)} \Phi$.

If $\mathcal{A}$ is such that a reduction theorem, e.g. from [82], applies to reduce the computation of the inner supremum in $\mathcal{U} \circ \Psi^{-1}$ to the supremum over convex combinations of Dirac masses, and the admissible set $\mathfrak{Q}$ is such that a reduction theorem applies to the computation of the outer supremum of $\sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}]$, then the identity (33) represents a nesting of reductions.

Let us now establish a result like (33). To do so will require addressing three questions: (1) What kind of function is $\mathcal{U} \circ \Psi^{-1}$? (2) What kind of measures $\mathbb{Q} \in \mathcal{M}(\mathcal{Q})$ can define an integral of a function with properties discovered from the answer to (1)? (3) Can we obtain a measurable solution operator to the optimization problem $(\mathcal{U} \circ \Psi^{-1})(q)$, where $q \in \mathcal{Q}$? To that end, let us first recall a definition of universally measurable functions.

Definition 3.9. Let $(T, \mathcal{T})$ be a measurable space, and for a positive measure ν on $(T, \mathcal{T})$, let $\mathcal{T}_\nu$ denote the ν-completion of $\mathcal{T}$. Let $\widehat{\mathcal{T}} := \bigcap_\nu \mathcal{T}_\nu$, where the intersection is over all positive bounded measures ν, denote the universally measurable sets. A $\widehat{\mathcal{T}}$-measurable function is said to be *universally measurable*.

At the heart of the commutative representation used for the nested reduction is the following optimal measurable selection lemma answering questions (1) and (3) above:

Lemma 3.10. *Let $\mathcal{A}$ be a Suslin space, let $\mathcal{Q}$ be a separable and metrizable space, and let $\Psi\colon \mathcal{A} \to \mathcal{Q}$ be measurable. Then, for any subset $T \subseteq \Psi(\mathcal{A})$,*

1. *$\mathcal{U} \circ \Psi^{-1}$ is $\widehat{\mathcal{B}}(T)$-measurable*
2. *for all $\delta > 0$, there exists a δ-suboptimal $\widehat{\mathcal{B}}(T)$-measurable section of Ψ; that is, a $\widehat{\mathcal{B}}(T)$-measurable function $\psi\colon T \to \mathcal{A}$ such that $\Psi(\psi(q)) = q$ for all $q \in T$ and*

$$\Phi\big(\psi(q)\big) > \mathcal{U}\big(\Psi^{-1}(q)\big) - \delta, \quad \text{for all } q \in T.$$

To answer question (2) above, define a *support* $\operatorname{supp}(\mathbb{Q})$ of a measure $\mathbb{Q} \in \mathcal{M}(\mathcal{Q})$, as in [4, Ch. 12.3], to be a closed set such that

- $\mathbb{Q}(\mathcal{Q} \setminus \operatorname{supp}(\mathbb{Q})) = 0$, and
- if $G \subseteq \mathcal{Q}$ is open and $G \cap \operatorname{supp}(\mathbb{Q}) \neq \varnothing$, then $\mathbb{Q}(G \cap \operatorname{supp}(\mathbb{Q})) > 0$.

When $\mathcal{Q}$ is a separable and metrizable space, it follows that it is second countable and therefore, by [4, Thm. 12.14], all $\mathcal{Q} \in \mathcal{M}(\mathcal{Q})$ have a uniquely defined support. Now consider a measure $\mathbb{Q} \in \mathcal{M}(\mathcal{Q})$ such that $\operatorname{supp}(\mathbb{Q}) \subseteq \Psi(\mathcal{A})$. Then, by Lemma 3.10, $\mathcal{U} \circ \Psi^{-1}$ is $\widehat{\mathcal{B}}(\operatorname{supp} \mathbb{Q})$-measurable. Therefore, the expected value $\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}]$ can be defined by integration with respect to the completion $\widehat{\mathbb{Q}}$:

$$\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] := \mathbb{E}_{\widehat{\mathbb{Q}}}[\mathcal{U} \circ \Psi^{-1}]. \tag{34}$$

More generally, for any universally measurable function f and any finite measure $\mathbb{Q}$, we define the expected value $\mathbb{E}_{\mathbb{Q}}[f]$ of f by

$$\mathbb{E}_{\mathbb{Q}}[f] := \mathbb{E}_{\widehat{\mathbb{Q}}}[f]. \tag{35}$$

Such a method of defining integrals of, possibly non-Borel measurable, but universally measurable, functions brings up many questions such as: when is it uniquely defined?; for a fixed integrand, when is the expectation operation affine in the measure?; does it have a change a variables formula? All such questions have nice answers and, although we are sure that this is classical, we cannot find a reference for these facts so we have included statements and proofs of the facts needed in this paper in Appendix A.1.

We now state our nested reduction theorem of the form (33):

Theorem 3.11. *Let $\mathcal{A}$ be a Suslin space, let $\mathcal{Q}$ be a separable and metrizable space, and let $\Psi\colon \mathcal{A} \to \mathcal{Q}$ measurable. Moreover, let $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ be such that $\operatorname{supp}(\mathbb{Q}) \subseteq \Psi(\mathcal{A})$ for all $\mathbb{Q} \in \mathfrak{Q}$. Then, for each $\mathbb{Q} \in \mathfrak{Q}$, $\Psi^{-1}\mathbb{Q}$ is non-empty. Moreover, the upper bound $\mathcal{U}(\Psi^{-1}\mathfrak{Q})$, defined in (20), satisfies*

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) = \sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}\big[\mathcal{U} \circ \Psi^{-1}\big]. \tag{36}$$

where the expectations on the right-hand side are defined as in (34). Finally, the expectation operator on the right-hand side is measure affine in $\mathbb{Q}$.

Remark 3.12. Note that (36) can be written

$$\sup_{\pi\in\Psi^{-1}\mathfrak{Q}} \mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\big] = \sup_{\mathbb{Q}\in\mathfrak{Q}} \Big[\mathbb{E}_{q\sim\mathbb{Q}}\big[\sup_{\mu\in\Psi^{-1}(q)} \Phi(\mu)\big]\Big]. \tag{37}$$

Remark 3.13. Since the right-hand side is measure affine in $\mathbb{Q}$, if $\mathbb{Q}$ is specified through (multi-)linear generalized moment inequalities, then the reduction theorems of [82] can be applied to obtain the supremum over $\mathbb{Q}$ by reducing $\mathbb{Q}$ to a convex combination of a finite number of Dirac masses on $\mathcal{Q}$. Moreover, if $\mathfrak{Q}$ consists of a single element, i.e. $\mathfrak{Q} = \{\mathbb{Q}\}$, then

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) = \mathcal{U}(\Psi^{-1}\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\big[\mathcal{U}\circ\Psi^{-1}\big], \tag{38}$$

and the right hand-side of (38) can be approximately evaluated via Monte Carlo sampling of $q \in \mathcal{Q}$ according to the measure $\mathbb{Q}$.

Remark 3.14. A similar theorem can obtained for the optimal lower bound $\mathcal{L}(\Psi^{-1}\mathfrak{Q})$. Throughout this paper, results given for optimal upper bounds $\mathcal{U}$ can be translated into results for optimal lower bounds $\mathcal{L}$ by considering the negative quantity of interest $-\Phi$ and for the sake of concision we will not write those results unless necessary.

Example 3.15. Consider again Example 3.7, where $\mathcal{X} := [0,1]$, $\mathcal{Q} = \mathbb{R}$, the admissible set $\mathcal{A} := \mathcal{M}([0,1])$, the quantity of interest $\Phi(\mu) := \mu[X \geq a]$ for some $a \in (0,1)$, the map $\Psi\colon \mathcal{A} \to \mathbb{R}$ is defined by $\Psi(\mu) := \mathbb{E}_\mu[X]$, and the set of admissible priors π on $\mathcal{A}$ is the collection

$$\Pi := \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_{\mu\sim\pi}\left[\mathbb{E}_\mu[X]\right] = q\}.$$

for some fixed $q \in (0,a)$. We will now demonstrate how the result $\mathcal{U}(\Pi) = q/a$ of (32) obtained by the primary reduction follows from the nested reduction theorem. To that end, observe that since $\Psi(\mathcal{A}) = [0,1] \subseteq \mathbb{R}$, by restricting to measures $\mathbb{Q} \in \mathcal{M}(\mathbb{R})$ with support $\mathrm{supp}(\mathbb{Q}) \subseteq [0,1]$, Theorem 3.11 implies that

$$\mathcal{U}(\Pi) = \sup_{\mathbb{Q}\in\mathfrak{Q}} \mathbb{E}_{q'\sim\mathbb{Q}}\left[\sup_{\mu\in\mathcal{M}([0,1])\,:\,\mathbb{E}_\mu[X]=q'} \mu[X\geq a]\right], \tag{39}$$

where $\mathfrak{Q}$ is the set of probability measures $\mathbb{Q}$ on $\mathbb{R}$ with support contained in $[0,1]$ such that $\mathbb{E}_{\mathbb{Q}}[Q] = q$. Theorem 4.1 of [82] shows that the inner supremum of $\mu[X \geq a]$ can be achieved by assuming that μ is the weighted sum of two Dirac masses, i.e.

$$\sup_{\substack{\mu\in\mathcal{M}([0,1])\\ \mathbb{E}_\mu[X]=q'}} \mu[X\geq a] = \sup_{\substack{\alpha,x_1,x_2\in[0,1]\\ \alpha x_1+(1-\alpha)x_2=q'}} (\alpha\delta_{x_1} + (1-\alpha)\delta_{x_2})[X\geq a]. \tag{40}$$

For $q' > a$, the supremum in the right-hand side of (40) is 1, and for $q' \leq a$, the supremum in the right-hand side of (40) is achieved by $x_2 = 0$, $x_1 = a$ and $\alpha = q'/a$, and so we conclude that

$$\sup_{\substack{\mu\in\mathcal{M}([0,1])\\ \mathbb{E}_\mu[X]=q'}} \mu[X\geq a] = \min\{1, \tfrac{q'}{a}\}.$$

Hence, by identifying the measures $\mathbb{Q} \in \mathcal{M}(\mathbb{R})$ with support $\mathrm{supp}(\mathbb{Q}) \subseteq [0,1]$ with $\mathcal{M}([0,1])$ in the obvious way, (39) becomes

$$\mathcal{U}(\Pi) = \sup_{\substack{\mathbb{Q}\in\mathcal{M}([0,1]) \\ \mathbb{E}_{\mathbb{Q}}[Q]=q}} \mathbb{E}_{q'\sim\mathbb{Q}}\left[\min\{1, \tfrac{q'}{a}\}\right]. \tag{41}$$

Using [82, Thm. 4.1] again, we obtain that the supremum in $\mathbb{Q}$ in the right-hand side of (41) is equal to the supremum over $\alpha, q_1, q_2 \in [0,1]$, of

$$\alpha \min\{1, \tfrac{q_1}{a}\} + (1-\alpha)\min\{1, \tfrac{q_2}{a}\} \tag{42}$$

subject to the constraint that $\alpha q_1 + (1-\alpha)q_2 = q$. This supremum is achieved by $q_1 = a$, $q_2 = 0$ and $\alpha = \frac{q}{a}$, and so we obtain that $\mathcal{U}(\Pi) = q/a$, in agreement with (32).

4. Optimal bounds on the posterior value

What happens to the optimal bounds (20) and (21) on the prior value $\mathbb{E}_\pi[\Phi]$, investigated in Section 3, after conditioning on the data? Does the interval corresponding to these optimal bounds shrink down to a single point as more and more data comes in? Does this interval shrink as the measurement noise on the data is reduced? What happens to posterior estimates associated with two distinct but close priors, possibly sharing the same marginal distribution on a high dimensional space? These are the questions that will be investigated in this section. Our answers will show that: (1) optimal bounds on posterior estimates *grow* as data comes in; (2) optimal bounds on posterior estimates *grow* as measurement noise is reduced (3) two priors sharing the same high-dimensional marginals can lead to *diametrically opposed* posterior estimates. In some sense these results can be seen as extreme occurrences of the dilation property observed in robust Bayesian inference [103].

As discussed in Section 2.4, let us now consider the case where the probability distribution of the data is a known function $\mathbb{D}(\mu)$ of the admissible candidates $\mu \in \mathcal{A}$. As shown in Section 2, directly conditioning measures $\pi \odot \mathbb{D}$ with respect to the random variable D representing the observed sample data would require manipulating regular conditional probabilities on $\mathcal{A} \times \mathcal{D}$.

Furthermore, in Bayesian statistics a prior π may represent a "subjective belief" about reality and, in such situations, the data may be sampled from $\pi^\dagger \cdot \mathbb{D}^\dagger$ which may be distinct from $\pi \cdot \mathbb{D}$. In frequentist analyses of Bayesian statistics $\pi^\dagger$ is called the "true" prior, or "data-generating distribution", and π a "subjective" prior (see [14] and references therein). Although it is known that the subjective prior π might be distinct from the true prior $\pi^\dagger$, one may still try to evaluate the conditional expectation of the quantity of interest Φ using π as the distribution on $\mathcal{A}$. We will show here that although the observation of the sample data d does not uniquely determine the true prior $\pi^\dagger$, it does determine a random subset of $\mathcal{M}(\mathcal{A})$ (i.e. a random subset of priors) denoted $\mathcal{R}(d)$ such that, $\pi^\dagger$-a.s., $\pi^\dagger \in \mathcal{R}(d)$. This observation is based on the following fundamental lemma:

Lemma 4.1. *For a strongly Lindelöf space $\mathcal{Y}$ and a Borel measure ν on $\mathcal{B}(\mathcal{Y})$, define*

$$E := \left\{ y \in \mathcal{Y} \;\middle|\; \begin{array}{c} \text{there is an open neighborhood } \mathcal{O}_y \\ \text{of } y \text{ such that } \nu(\mathcal{O}_y) = 0 \end{array} \right\}.$$

Then $\nu(E) = 0$.

Remark 4.2. Recall that a Lindelöf space is a topological space such that any open cover has a countable subcover and a strongly Lindelöf space is such that any open subset is Lindelöf. Since $\mathcal{D}$ is assumed to be Suslin from Section 2.4, and Suslin implies strongly Lindelöf, Lemma 4.1 shows that any open neighborhood B_d of any observed value $d \in \mathcal{D}$ has nonzero measure with probability 1.

Remark 4.3. Any separable Hilbert space, in particular the Euclidian space $\mathbb{R}^k$, is strongly Lindelöf. In this situation, Lemma 4.1 implies that if for any observation y generated by a law $\nu \in \mathcal{M}(\mathcal{Y})$ we place an open ball $B_{r(y)}(y)$ of non-zero radius $r(y) > 0$ about y, then with ν-probability 1 we have $\nu(B_{r(y)}(y)) > 0$. That is,

$$\nu\big(\{y \in \mathcal{Y} \mid \nu(B_{r(y)}(y)) > 0\}\big) = 1.$$

Now suppose the data d are generated according to a probability measure $\pi^\dagger \cdot \mathbb{D}$ (where $\pi^\dagger$ is the "true" prior). We conclude from Lemma 4.1 that when we observe a sample d, if we assume that $\pi^\dagger \in \mathcal{R}(d)$ where

$$\mathcal{R}(d) := \big\{\pi \in \mathcal{M}(\mathcal{A}) \,\big|\, \pi \cdot \mathbb{D}[B] > 0 \text{ for all } B \text{ open containing } d\big\},$$

then we will be correct in this assumption with $\pi^\dagger \cdot \mathbb{D}$-probability 1. Therefore, when the data d are generated and we observe that $d \in B_d$ where B_d is an open subset containing the data d (to keep our notation simple, we will, later on, drop d in the notation B_d), then we restrict our attention to priors $\pi \in \Pi$ such that $\pi \cdot \mathbb{D}[B_d] > 0$. That is to say, we restrict our attention to the intersection of Π with the set of priors π such that $\pi \in \mathcal{M}(\mathcal{A})$ and $\pi \cdot \mathbb{D}[B_d] > 0$. We write Π_{B_d} for this intersection, i.e.

$$\Pi_{B_d} := \{\pi \in \Pi \,|\, \pi \cdot \mathbb{D}[B_d] > 0\}.$$

If Π_{B_d} is void, then we assert that "$\pi^\dagger$ is not contained in Π" and we know that this assertion is true with $\pi^\dagger \cdot \mathbb{D}$-probability 1 on the realization of the data d. Conversely, if $\pi^\dagger$ is contained in Π, then Π_{B_d} must, with $\pi^\dagger \cdot \mathbb{D}$-probability 1 on the realization of the data d, still contain $\pi^\dagger$ (in particular it must be non-empty).

Happily, this approach also facilitates the efficient computation of the conditional expectations because now they have a simple representation. Indeed, consider the conditional expectation of an object of interest Φ given a prior π and data map $\mathbb{D}$, conditioned on a subset $B \in \mathcal{B}(\mathcal{D})$ such that $\pi \cdot \mathbb{D}[B] > 0$. It follows from (17) and (18) that the conditional expectation of Φ given B is

$$\mathbb{E}_{\pi \odot \mathbb{D}}\big[\Phi \big| B\big] := \frac{\mathbb{E}_{(\mu,d) \sim \pi \odot \mathbb{D}}\big[\Phi(\mu) \mathbb{1}_B(d)\big]}{\pi \cdot \mathbb{D}[B]},$$

which, using (17) and (18), leads to

$$\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big] = \frac{\mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\mathbb{D}(\mu)[B]\big]}{\mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big]}. \tag{43}$$

Moreover, recall that this conditional expectation is the best mean squared approximation of Φ under the measure $\pi\odot\mathbb{D}$, given the information that $D\in B$, i.e.

$$\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big] = \arg\min_{m\in\mathbb{R}} \mathbb{E}_{\pi\odot\mathbb{D}}\Big[\big(\Phi-m\big)^2\Big|B\Big]. \tag{44}$$

Consequently, for any open subset $B\subseteq\mathcal{D}$, we define

$$\Pi_B := \{\pi\in\Pi\,|\,(\pi\cdot\mathbb{D})[B]>0\}. \tag{45}$$

where, by (18),

$$\pi\cdot\mathbb{D}[B] := \mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big]. \tag{46}$$

Then, since $(\pi\cdot\mathbb{D})[B]>0$, the formula (43) for conditional expectation implies that

$$\mathcal{U}(\Pi|B) := \sup_{\pi\in\Pi_B} \mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big] \tag{47}$$

$$\mathcal{L}(\Pi|B) := \inf_{\pi\in\Pi_B} \mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big] \tag{48}$$

where

$$\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big] = \frac{\mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\mathbb{D}(\mu)[B]\big]}{\mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big]}. \tag{49}$$

Finally, if B is an open neighborhood containing the sample data d, then it follows that $\mathcal{U}(\Pi|B)$ and $\mathcal{L}(\Pi|B)$ are optimal upper and lower bounds on the posterior values $\mathbb{E}_{\pi\odot\mathbb{D}}[\Phi|B]$, given the observation $D\in B$, over all $\pi\in\Pi$ such that $\pi\cdot\mathbb{D}[B]>0$.

Example 4.4. When Φ is the indicator function of the set $\{\mu\mid\mu[X\geq a]>\epsilon\}$ (i.e. the set of unsafe μ), $\mathcal{U}(\Pi|B)$ and $\mathcal{L}(\Pi|B)$ are optimal upper and lower bounds on the "posterior probability" that the system is unsafe given the observation $D\in B$ (and the set Π of priors and observation maps respectively).

4.1. General information bounds on posterior values

Now let $B\subseteq\mathcal{D}$ be open and let

$$\mathcal{A}_{\Pi_B} := \{\mu\in\mathcal{A}\,|\,\delta_\mu\in\Pi \text{ and } \mathbb{D}(\mu)[B]>0\}, \tag{50}$$

$$\mathcal{U}(\mathcal{A}_{\Pi_B}) := \sup_{\mu\in\mathcal{A}_{\Pi_B}} \Phi(\mu),$$

and use $\mathcal{L}$ for the corresponding infimum. The following theorem is a straightforward consequence of (43):

Theorem 4.5. *It holds true that*

$$\mathcal{U}(\mathcal{A}_{\Pi_B}) \leq \mathcal{U}(\Pi_B) \leq \mathcal{U}(\mathcal{A}),$$

and

$$\mathcal{L}(\mathcal{A}) \leq \mathcal{L}(\Pi_B) \leq \mathcal{L}(\mathcal{A}_{\Pi_B}).$$

Moreover, if $\mathcal{A}_{\Pi_B}$ is non empty, then

$$\mathcal{L}(\mathcal{A}) \leq \mathcal{L}(\Pi_B) \leq \mathcal{L}(\mathcal{A}_{\Pi_B}) \leq \mathcal{U}(\mathcal{A}_{\Pi_B}) \leq \mathcal{U}(\Pi_B) \leq \mathcal{U}(\mathcal{A}).$$

Remark 4.6. The dependence of $\mathcal{U}(\mathcal{A}_{\Pi_B})$ and $\mathcal{L}(\mathcal{A}_{\Pi_B})$ on the sample data is very weak. In particular, if $\mathbb{D}$ corresponds to observing i.i.d. realizations of $(X + \xi, f^\dagger(X) + \xi')$ where ξ and ξ' are centered Gaussian random variables of arbitrarily small (non zero) variance, then it can be shown that $\mathcal{U}(\mathcal{A}_{\Pi_B}) = \mathcal{U}(\mathcal{A}_\Pi)$ and $\mathcal{L}(\mathcal{A}_{\Pi_B}) = \mathcal{L}(\mathcal{A}_\Pi)$. In that situation, if $\mathcal{L}(\mathcal{A}_\Pi) < \mathcal{U}(\mathcal{A}_\Pi)$, then $\mathcal{U}(\mathcal{A}_{\Pi_B}) - \mathcal{L}(\mathcal{A}_{\Pi_B})$ remains bounded away from 0 by a strictly positive constant that is independent of $\mathfrak{D}$ and B, which, in particular, implies that the range of achievable posterior values cannot shrink towards $\Phi(\mu^\dagger)$ regardless of the number of observed i.i.d. samples. The presence of such information bounds suggests that the consistency of Bayesian estimators cannot be established independently of (uniformly in) the choice of priors (this point will also be substantiated by Theorem 4.13).

4.2. Primary reduction for posterior values

As in Section 3.2.1, when priors are specified through finite-dimensional inequalities, it is possible to provide a reduction of the computation of $\mathcal{U}(\Pi|B)$ on the primary space. To that end, let $\mathcal{M}_+(\mathcal{A})$ denote the set of positive bounded measures on $\mathcal{A}$ and let us extend the "expectation notation" to mean integration with respect to a positive measure in the natural way: for a measurable function ψ and a $\pi_+ \in \mathcal{M}_+(\mathcal{A})$ define

$$\mathbb{E}_{\pi_+}[\psi] := \int_{\mathcal{A}} \psi \, \mathrm{d}\pi_+$$

if the integral exists.

Let $\psi_0, \ldots, \psi_n$ be real-valued measurable functions on $\mathcal{A}$ and define

$$\Pi_+ := \left\{ \pi_+ \in \mathcal{M}_+(\mathcal{A}) \,\middle|\, \mathbb{E}_{\pi_+}[\psi_0] = 1, \text{ and } \mathbb{E}_{\pi_+}[\psi_i] = 0 \text{ for } i = 1, \ldots, n \right\},$$

where implicit in the definition is that all $n+1$ integrals exist, and let

$$\Pi_{+,n} := \Pi_+ \cap \Delta(n)$$

be the set of those measures in Π_+ that are non-negative sums of $n+1$ Dirac masses. The following theorem is a generalization of [82, Thm. 4.1] to positive measures (see also [107, Thm. 3.2] from which the proof of [82, Thm. 4.1] was derived).

Theorem 4.7. *If $\mathcal{A}$ is a Suslin space, then*

$$\sup_{\pi_+\in\Pi_+}\mathbb{E}_{\pi_+}[\Phi]=\sup_{\pi_+\in\Pi_{+,n+1}}\mathbb{E}_{\pi_+}[\Phi]. \tag{51}$$

Furthermore, if ψ_0 is non-negative on $\mathcal{A}$ and there exists a measurable function φ such that $\Phi=\psi_0\varphi$, then

$$\sup_{\pi_+\in\Pi_+}\mathbb{E}_{\pi_+}[\Phi]=\sup_{\pi_+\in\Pi_{+,n}}\mathbb{E}_{\pi_+}[\Phi]. \tag{52}$$

Theorem 4.7 can be used to produce a primary reduction of $\mathcal{U}(\Pi\odot_B\mathcal{D})$ when Π is defined by a finite number of equalities. To state the theorem, recall that, for arbitrary Π and B, the definition

$$\Pi_B:=\{\pi\in\Pi\,|\,\pi\cdot\mathbb{D}[B]>0\}$$

of (45), where by (46)

$$\pi\cdot\mathbb{D}[B]:=\mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big];$$

recall also the notation of (47)

$$\mathcal{U}(\Pi|B):=\sup_{\pi\in\Pi_B}\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big];$$

and recall the result (43) that, for any $\pi\in\Pi_B$,

$$\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big]=\frac{\mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\mathbb{D}(\mu)[B]\big]}{\mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big]}.$$

The proof of the following theorem is obtained by first proving the theorem for equality constraints $Z=\{q\}$, by observing that $\mathcal{U}(\Pi(q)|B)$ is a linear fractional optimization problem in π and utilizing the fact that such problems are equivalent to linear problems [27], and then applying Theorem 4.7. To extend the result to the subset $Z\subseteq\mathbb{R}^n$, one uses a layercake approach as in the proof of Theorem 3.6. As in Section 3, the following primary reduction theorem, Theorem 4.8, will be formulated in canonical form and the nested reduction theorem, Theorem 4.11, will be in the general form.

Theorem 4.8. *Let $\mathcal{A}$ be Suslin and let $\Psi\colon\mathcal{A}\to\mathbb{R}^n$ be measurable. For $Z\subseteq\mathbb{R}^n$, let $\Pi(Z):=\{\pi\in\mathcal{M}(\mathcal{A})\mid\mathbb{E}_\pi[\Psi]\in Z\}$. Then $\mathcal{U}(\Pi(Z)|B)$ is equal to the supremum over $\alpha_i\geq 0$, $q\in Z$ and $\mu_i\in\mathcal{A}$ of*

$$\sum_{i=0}^{n}\alpha_i\Phi(\mu_i)\mathbb{D}(\mu_i)[B]$$

subject to the constraints

$$\sum_{i=0}^{n}\alpha_i\big(\Psi(\mu_i)-q\big)=0$$

and

$$\sum_{i=0}^{n} \alpha_i \mathbb{D}(\mu_i)[B] = 1. \tag{53}$$

Example 4.9. Consider again Example 3.7 with the admissible set $\mathcal{A} := \mathcal{M}([0,1])$, the quantity of interest $\Phi(\mu) := \mu[X \geq a]$, the map $\Psi(\mu) := \mathbb{E}_\mu[X]$ and the set of admissible priors

$$\Pi := \left\{ \pi \in \mathcal{M}(\mathcal{A}) \,\middle|\, \mathbb{E}_{\mu\sim\pi}\left[\mathbb{E}_\mu[X]\right] = q \right\}.$$

for some $q \in (0,a)$. We saw in Example 3.7 that $\mathcal{U}(\Pi) = \frac{q}{a}$. Now suppose that we observe the random variable $D := (X_1, \ldots, X_n)$ corresponding to n i.i.d. samples of $\mu^\dagger \in \mathcal{A}$. More precisely, we observe $D \in B$ where $B = B_1 \times \cdots B_n$ and B_i is the ball in $(0,1)$ of center x_i and radius ρ, $x_i \in (0,1)$ and $0 < \rho \ll 1/n$. Let $\mathbb{D}^n$ denote the data map corresponding to taking n i.i.d. samples, that is, $\mathbb{D}^n(\mu) := \mu \otimes \cdots \otimes \mu$, and observe that $\mathbb{D}^n(\mu)[B] = \prod_{i=1}^n \mu[B_i]$.

Theorem 4.8 implies that $\mathcal{U}(\Pi \odot_B \mathbb{D}^n)$ is equal to the supremum over $\alpha_1, \alpha_2 \geq 0$, $\mu_1, \mu_2 \in \mathcal{A}$ of

$$\alpha_1 \mu_1[X \geq a]\mathbb{D}^n(\mu_1)[B] + \alpha_2 \mu_2[X \geq a]\mathbb{D}^n(\mu_2)[B]$$

subject to the constraints

$$\alpha_1(\mathbb{E}_{\mu_1}[X] - q) + \alpha_2(\mathbb{E}_{\mu_2}[X] - q) = 0,$$

$$\alpha_1 \mathbb{D}^n(\mu_1)[B] + \alpha_2 \mathbb{D}^n(\mu_2)[B] = 1,$$

with $\mathbb{D}^n(\mu)[B] = \prod_{i=1}^n \mu(B_i)$. Introducing slack variables $\beta_{1,i} := \mu_1[B_i]$ and $\beta_{2,i} := \mu_2[B_i]$ as n linear constraints on μ_1 and n linear constraints on μ_2 we obtain (from [82, Thm. 4.1]) that the supremum can be achieved by assuming that each μ_i is the weighted sum of at most $n+2$ Dirac masses. Assuming that the B_i are non intersecting balls of radius $\rho \ll 1/n$ centered on $x_1, \ldots, x_n$, n of these Dirac masses will have to be put at $x_1, \ldots, x_n$; for optimality, the two others will have to be put at 0 and a (with weights p_1 and p_2). Introducing $\gamma_1 = \alpha_1 \mathbb{D}^n(\mu_1)[B]$ and $\gamma_2 = \alpha_2 \mathbb{D}^n(\mu_2)[B]$, it follows that $\mathcal{U}(\Pi \odot_B \mathbb{D}^n)$ is equal (as $\rho \downarrow 0$) to the supremum over $\gamma_1, \gamma_2 \geq 0$, $p_1, p_2 \in [0,1]$ of

$$\gamma_1 p_1 + \gamma_2 p_2$$

subject to the constraints

$$\gamma_1 + \gamma_2 = 1,$$

and

$$\gamma_1 \frac{(ap_1 + \sum_{i=1}^n x_i \beta_{1,i}) - q}{\prod_{i=1}^n \beta_{1,i}} + \gamma_2 \frac{(ap_2 + \sum_{i=1}^n x_i \beta_{2,i}) - q}{\prod_{i=1}^n \beta_{2,i}} = 0.$$

By considering $0 < \beta_{i,j} \ll 1$ it is easy to obtain that $\mathcal{U}(\Pi \odot_B \mathbb{D}^n) = 1$.

Example 4.10. We will now use Theorem 4.8 to prove equation (8) of Subsection 1.4. Let Φ be defined as in Subsection 1.4. Let $\mathcal{A}(\alpha)$ and $\Pi(\alpha)$ be defined as in (6) and (7). Then, Theorem 4.8 implies that $\mathcal{U}(\Pi(\alpha)|B_\delta^n)$, the least upper bound on posterior values, is equal to the supremum over $\alpha_1, \alpha_2 \geq 0$, $\mu_1, \mu_2 \in \mathcal{A}(\alpha)$ of

$$\alpha_1 \mu_1[X \geq a]\mu_1^n[B_\delta^n] + \alpha_2 \mu_2[X \geq a]\mu_2^n[B_\delta^n]$$

subject to the constraints

$$\begin{cases} \alpha_1(\mathbb{E}_{\mu_1}[X] - m) + \alpha_2(\mathbb{E}_{\mu_2}[X] - m) = 0, \\ \alpha_1 \mu_1^n[B_\delta^n] + \alpha_2 \mu_2^n[B_\delta^n] = 1, \end{cases}$$

where we have used the notation $\mu^n[B_\delta^n] := \prod_{i=1}^n \mu(B_\delta(x_i))$.

Introducing $\gamma_1 = \alpha_1 \mu_1^n[B_\delta^n]$ and $\gamma_2 = \alpha_2 \mu_2^n[B_\delta^n]$, it follows that $\mathcal{U}(\Pi(\alpha)|B_\delta^n)$ is equal to the supremum over $\gamma_1, \gamma_2 \geq 0$, $\mu_1, \mu_2 \in \mathcal{A}(\alpha)$ of

$$\gamma_1 \mu_1[X \geq a] + \gamma_2 \mu_2[X \geq a]$$

subject to the constraints

$$\begin{cases} \gamma_1 + \gamma_2 = 1, \\ \gamma_1 \frac{\mathbb{E}_{\mu_1}[X]-m}{\mu_1^n[B_\delta^n]} + \gamma_2 \frac{\mathbb{E}_{\mu_2}[X]-m}{\mu_2^n[B_\delta^n]} = 0. \end{cases}$$

which can be simplified to the supremum over $\mu_1, \mu_2 \in \mathcal{A}(\alpha)$ of

$$\frac{1}{1 + \frac{\mathbb{E}_{\mu_1}[X]-m}{m-\mathbb{E}_{\mu_2}[X]} \frac{\mu_2^n[B_\delta^n]}{\mu_1^n[B_\delta^n]}} \mu_1[X \geq a] + (1 - \frac{1}{1 + \frac{\mathbb{E}_{\mu_1}[X]-m}{m-\mathbb{E}_{\mu_2}[X]} \frac{\mu_2^n[B_\delta^n]}{\mu_1^n[B_\delta^n]}}) \mu_2[X \geq a] \tag{54}$$

By introducing slack variables for $m_1 = \mathbb{E}_{\mu_1}[X]$ and $m_2 = \mathbb{E}_{\mu_2}[X]$, maximizing (54) with m_1 and m_2, then taking a supremum over m_1, m_2, one obtains that the supremum of (54) is achieved, in the limit $\delta \downarrow 0$, in the configuration where μ_1 puts most of its mass on a, μ_2 puts most of its mass on 0, and $\frac{\mu_2^n[B_\delta^n]}{\mu_1^n[B_\delta^n]} \approx \frac{1}{\alpha^2}$ which yields

$$\lim_{\delta \to 0} \mathcal{U}\big(\Pi(\alpha)|B_\delta^n\big) = \frac{1}{1 + \frac{1}{\alpha^2}\frac{a-m}{m}}. \tag{55}$$

4.3. Nested reduction for posterior values

Here, as in Section 3.2, we show how the optimization problems (47) and (48) can be reduced to nested OUQ optimization problems (i.e. nested problems analogous to (14) and (15)) when the collection Π of admissible priors is defined by how they push forward by a measurable mapping $\Psi\colon \mathcal{A} \to \mathcal{Q}$. That is, we specify a feature space $\mathcal{Q}$, a measurable map $\Psi\colon \mathcal{A} \to \mathcal{Q}$, a subset $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ and define the admissible set of priors by

$$\Pi := \Psi^{-1}\mathfrak{Q} = \{\pi \in \mathcal{M}(\mathcal{A}) \mid \Psi\pi \in \mathfrak{Q}\}.$$

As before, we focus on reducing the upper bound

$$\mathcal{U}\big(\Psi^{-1}\mathfrak{Q}\big|B\big) := \sup_{\pi\in(\Psi^{-1}\mathfrak{Q})_B} \mathbb{E}_{\pi\odot\mathbb{D}}[\Phi|B]. \tag{56}$$

Theorem 4.11. *Let $\mathcal{A}$ be a Suslin space, let $\mathcal{Q}$ be a separable and metrizable space, and let $\Psi\colon\mathcal{A}\to\mathcal{Q}$ be measurable. Moreover, let $\mathfrak{Q}\subseteq\mathcal{M}(\mathcal{Q})$ be such that $\operatorname{supp}(\mathbb{Q})\subseteq\Psi(\mathcal{A})$ for all $\mathbb{Q}\in\mathfrak{Q}$. Then, for each $\mathbb{Q}\in\mathfrak{Q}$, $\Psi^{-1}\mathbb{Q}$ is non-empty. Moreover, the upper bound $\mathcal{U}(\Psi^{-1}\mathfrak{Q}|B)$, defined in (56), satisfies*

$$\mathcal{U}\big(\Psi^{-1}\mathfrak{Q}\big|B\big) = \sup\left\{\lambda\in\mathbb{R}\,\middle|\,\sup_{\mathbb{Q}\in\mathfrak{Q}}\mathbb{E}_{q\sim\mathbb{Q}}\left[\sup_{\mu\in\Psi^{-1}(q)}\big(\Phi(\mu)-\lambda\big)\mathbb{D}(\mu)[B]\right]>0\right\}, \tag{57}$$

where the expectations on the right-hand side are defined as in (35). Finally, the expectation operator on the right-hand side is measure affine in $\mathbb{Q}$, as defined in (3.3).

Remark 4.12. Note that Theorem 4.11 is more general than Theorem 4.8 because its application does not require the assumption that $\Psi^{-1}\mathfrak{Q}$ is defined via generalized moments constraints.

The following theorem is our main result. It shows not only that the right-hand side of the assertion (57) of Theorem 4.11 depends on the sample data in a very weak way, but also that, under very mild assumptions, the observation of this sample data leads to an increase (rather than a decrease) of the least upper bound on the quantity of interest:

Theorem 4.13 (Main Brittleness Theorem)**.** *Let $\mathcal{A}$ be a Suslin space, let $\mathcal{Q}$ be a separable and metrizable space, and let $\Psi\colon\mathcal{A}\to\mathcal{Q}$ be measurable. Moreover, let $\mathfrak{Q}\subseteq\mathcal{M}(\mathcal{Q})$ be such that $\operatorname{supp}(\mathbb{Q})\subseteq\Psi(\mathcal{A})$ for all $\mathbb{Q}\in\mathfrak{Q}$. Suppose that, for all $\delta>0$, there exists some $\mathbb{Q}\in\mathfrak{Q}$ such that*

$$\mathbb{E}_{q\sim\mathbb{Q}}\left[\inf_{\mu\in\Psi^{-1}(q)}\mathbb{D}(\mu)[B]\right]=0 \tag{58}$$

and

$$\mathbb{P}_{q\sim\mathbb{Q}}\left[\sup_{\mu\in\Psi^{-1}(q),\,\mathbb{D}(\mu)[B]>0}\Phi(\mu) > \sup_{\mu\in\mathcal{A}}\Phi(\mu)-\delta\right]>0. \tag{59}$$

Then

$$\mathcal{U}\big(\Psi^{-1}\mathfrak{Q}\big|B\big)=\mathcal{U}(\mathcal{A}). \tag{60}$$

Remark 4.14. Note that the convention that $\sup\varnothing=-\infty$ implies that, if the assumption (59) is satisfied, then there is a measure $\mathbb{Q}\in\mathfrak{Q}$ such that the set of q such that $\mathbb{D}(\mu)[B]>0$ for some $\mu\in\Psi^{-1}(q)$ has strictly positive $\mathbb{Q}$-measure.

Remark 4.15. Theorem 4.13 states that if there exists $\mathbb{Q}\in\mathfrak{Q}$ putting some mass on a neighborhood of the values q of Ψ where $\sup_{\mu\in\Psi^{-1}(q)}\Phi(\mu)$ achieves its supremum, then

$$\mathcal{U}\big(\Psi^{-1}(\mathfrak{Q})\big|B\big)=\mathcal{U}(\mathcal{A}).$$

On the other hand, Theorem 3.2 asserts that

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) \leq \mathcal{U}(\mathcal{A}), \tag{61}$$

so we conclude that

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) \leq \mathcal{U}\big(\Psi^{-1}\mathfrak{Q}\big|B\big). \tag{62}$$

That is, *observing the sample data does not improve the optimal bound!* Moreover, when the inequality (61) is strict, if we define

$$\delta := \mathcal{U}(\mathcal{A}) - \mathcal{U}(\Psi^{-1}\mathfrak{Q}) > 0$$

then it follows that

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) + \delta \leq \mathcal{U}\big(\Psi^{-1}(\mathfrak{Q})\big|B\big), \tag{63}$$

from which we conclude that when the inequality (61) is strict, *observing the sample data makes the optimal bound worse!* In other words, after the observation of the sample data (which may be limited to a single realization of X under the measure $\mu^\dagger$, or an arbitrary large number of independent samples of X_i) the optimal upper bound on the quantity of interest,

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) = \sup_{\pi\in\Psi^{-1}\mathfrak{Q}} \mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\big],$$

increases to

$$\mathcal{U}(\mathcal{A}) = \sup_{\mu\in\mathcal{A}} \Phi(\mu).$$

Example 4.16. Consider $\mathcal{A} := \mathcal{M}([0,1])$, $\Phi(\mu) = \mathbb{E}_\mu[X]$, $\mathbb{D}^n(\mu) := \mu\otimes\cdots\otimes\mu$. In this example are interested in estimating the mean of X under some unknown measure $\mu^\dagger \in \mathcal{A}$ and we observe $d = (d_1,\ldots,d_n)$, n i.i.d. samples from X; note that n can be very large. The sample data contain information on $\mu^\dagger$ through the fact that their distribution is $\mathbb{D}^n(\mu^\dagger) = \mu^\dagger\otimes\cdots\otimes\mu^\dagger$ (i.e. although the distribution of the sample data is unknown, its dependency structure, as a functional of $\mu^\dagger$, is known).

Let k be a (possibly large) number. Define Π to be the set of priors π under which the distribution of $(\mathbb{E}_\mu[X],\ldots,\mathbb{E}_\mu[X^k])$ is $\mathbb{Q}$, where $\mathbb{Q}$ is a distribution on $\mathbb{R}^k$ such that $\mathbb{E}_\mu[X]$ (its first marginal) is uniformly distributed on $[0,1]$ and such that the (conditional) distribution of $\mathbb{E}_\mu[X^2]$ conditioned on $\mathbb{E}_\mu[X] = q_1$ is the uniform distribution on the interval

$$\left[\inf_{\mu\in\mathcal{A},\,\mathbb{E}[X]=q_1} \mathbb{E}_\mu[X^2],\ \sup_{\mu\in\mathcal{A},\,\mathbb{E}[X]=q_1} \mathbb{E}_\mu[X^2]\right]$$

and such that the conditional distributions of the other marginals $\mathbb{E}_\mu[X^k]$ are defined iteratively in the same manner. For this example, note that $\Psi(\mu) = (\mathbb{E}_\mu[X],\ldots,\mathbb{E}_\mu[X^k])$. Note that, for $q := (q_1,\ldots,q^k)$ in the range of Ψ (i.e. $\Psi(\mathcal{A})$), $\Psi^{-1}(q)$ is the subset of measures $\mu \in \mathcal{M}([0,1])$ such that $\mathbb{E}_\mu[X^i] = q_i$ for

$1 \leq i \leq k$. Let B be defined as $B_1 \times \cdots B_n$ where each B_i is a ball of radius ρ containing d_i.

We will now use Theorem 4.13 to compute optimal bounds on the posterior values of $\Phi(\mu) = \mathbb{E}_\mu[X]$. We will focus our attention on the upper bound. First observe that in this example $\mathfrak{Q}$ is reduced to the single measure $\mathbb{Q}$ constructed above and $\mathfrak{D}$ is reduced to the single data map $\mathbb{D}^n$.

Let us first check that condition (59) is always satisfied (irrespective of the value of the data d). Note that condition (59) is satisfied if for all $\delta > 0$ there exists a subset of values of q of strictly positive $\mathbb{Q}$-measure such that $\{\mu \in \Psi^{-1}(q) \mid \mathbb{D}^n(\mu)[B] > 0 \text{ and } \mathbb{E}_\mu[X] \geq 1 - \delta\}$ is non empty. So, let $\delta > 0$ be arbitrary and define μ_d to be the empirical distribution of d, i.e.

$$\mu_d := \frac{\sum_{i=1}^n \delta_{d_i}}{n}.$$

Define

$$\mathcal{A}_\delta := \{\mu \in \mathcal{A} \mid \mathbb{E}_\mu[X] \geq 1 - \delta/2\}.$$

One can show by induction that $\Psi(\mathcal{A}_\delta)$ has a non-empty interior and that any open subset of $\Psi(\mathcal{A})$ has strictly positive $\mathbb{Q}$-measure. Let q^* be a point in the interior of $\Psi(\mathcal{A}_\delta)$, and let $B_\tau(q^*)$ be a ball of center q^* and radius τ such that $B_{2\tau}(q^*)$ is contained in the interior of $\Psi(\mathcal{A}_\delta)$. Note that $B_\tau(q^*)$ has strictly positive $\mathbb{Q}$-measure. Furthermore, for ϵ sufficiently small, for each $q \in B_\tau(q^*)$ there exists $q' \in B_{2\tau}(q^*)$ and $\mu \in \Psi^{-1}(q')$ such that $\mu_\epsilon := (1-\epsilon)\mu + \epsilon\mu_d \in \Psi^{-1}(q)$. Since $\mathbb{D}^n(\mu_\epsilon)[B] > 0$ and $\mathbb{E}_\mu[X] \geq 1 - \delta/2$, it follows that (59) is satisfied (irrespective of the value of the data d).

Let us now consider condition (58). Observe that condition (58) is satisfied if for $\mathbb{Q}$-almost all $q \in \Psi(\mathcal{A})$ and all $\epsilon > 0$, there exists $\mu \in \Psi^{-1}(q)$ such that $\mathbb{D}^n(\mu)[B] < \epsilon$. Assume that d contains at least $k+2$ distinct points and that ρ is strictly smaller than half of the minimal distance between two of such points, so that the associated B_i do not overlap; note that this assumption is satisfied with probability converging to one (as $n \to \infty$) if the data are sampled from a measure $\mu^\dagger$ that is absolutely continuous with respect to the Lebesgue measure on $[0,1]$. Let $q \in \Psi(\mathcal{A})$; by the reduction theorems of [82] there exists $\mu_q \in \Psi^{-1}(q)$ such that μ_q is the weighted sum of at most $k+1$ Dirac masses in $[0,1]$. Since there exist at least $k+2$ non-overlapping B_i we have $\mathbb{D}^n(\mu_q)[B] = 0$ which implies condition (58). Hence, Theorem 4.13 implies that, for this (possibly) highly constrained problem characterized by a (possibly) large number of sampled data points, the optimal bounds on the posterior values of $\mathbb{E}_\mu[X]$ are zero and one whereas the set of prior values of $\mathbb{E}_\mu[X]$ is the single point $\{\frac{1}{2}\}$.

Remark 4.17. For a thorough analysis of Example 4.16 we refer to [80] where, in particular, a *quantitative* version of Theorem 4.13 is developed and then applied to Example 4.16. Curiously, a refined analysis of the integral geometry of the truncated Hausdorff moment space, used to demonstrate the approximate satisfaction of the conditions of Theorem 4.13, is shown in [80] to lead to a new family of Selberg integral formulas. See [46] for a discussion of their importance.

Remark 4.18. Note that the assumptions of Theorem 4.13 are extremely weak. In plain words, Theorem 4.13 implies that if the probability of observing the data can be arbitrary small under priors contained in $\mathcal{A}$ that are putting mass near the extreme values of Φ, then the optimal bounds on posterior values are the extreme values of Φ in $\mathcal{A}$ (even if the data comes in the form of a large number of samples and the set of priors is highly constrained). Example 4.16 illustrates that one consequence of Theorem 4.13 is that Bayesian posteriors are not robust, and in fact are fragile with respect to the choices of priors constrained by marginals, even with a highly constrained subset of priors of $\mathcal{M}(\mathcal{A})$.

Moreover, if Π is convex, then by considering priors of the form $\pi_0\lambda+(1-\lambda)\pi_1$ with $\pi_0, \pi_1 \in \Pi$, $\pi_0 \cdot \mathbb{D}[B] > 0$ and $\pi_1 \cdot \mathbb{D}[B] > 0$, it is easy to see that the Bayesian posterior can take any value in the interval $(\mathcal{L}(\mathcal{A}),\mathcal{U}(\mathcal{A}))$, irrespective of the data. In addition, it is easy to observe that even including the quantity of interest Φ in the marginal Ψ does not prevent this fragility. Theorem 4.13 also leads to the following apparent paradoxes when the Bayesian framework is applied to the space $\mathcal{A}$: (1) Posteriors with different priors may diverge as more and more data comes in; (2) When the sample data is observed with some (say Gaussian) measurement noise of variance σ^2, then, the optimal bound $\mathcal{U}(\Psi^{-1}(\mathfrak{Q})|B)$ on the quantity of interest Φ converges towards $\mathcal{U}(\Psi^{-1}(\mathfrak{Q}))$ as $\sigma^2 \to \infty$. That is, if one interprets optimal bounds on posterior values as uncertainty bounds, then one would reach the paradoxical conclusion that adding measurement uncertainty decreases the uncertainty of the quantity of interest. The idea of the proof of this assertion is based on the following observation:

Let y be the (noisy) measurement whose distribution given the value of the data d is assumed to be independent of μ. Write $p_\sigma(d)[B]$ for the probability that the value of y belongs to a set B and observe that the conditional value of the quantity of interest Φ given the $y \in B$ is equal to

$$\frac{\mathbb{E}_\pi\left[\Phi(\mu)\mathbb{E}_{d\sim\mathbb{D}(\mu)}\left[p_\sigma(d)[B]\right]\right]}{\mathbb{E}_\pi\left[\mathbb{E}_{d\sim\mathbb{D}(\mu)}\left[p_\sigma(d)[B]\right]\right]}. \tag{64}$$

We deduce that if $p_\sigma(d)[B]/p_\sigma(d')[B]$ converges towards one as the level of noise $\sigma \to \infty$ uniformly in $(d, d') \in [0, 1]^2$ (which is the case if the data in Example 4.9 is observed with Gaussian noise of increasing variance, see also Example 4.19 below), then (64) converges towards the prior value of Φ as $\sigma \to \infty$ uniformly in π.

The fact that optimal bounds on prior values may become less precise after conditioning is known as the *dilation phenomenon* in robust Bayesian inference [103], and, in some sense, the brittleness results presented in this paper could be seen as an extreme occurrence of this phenomenon.

Example 4.19. Consider again Example 4.9 with the set of admissible priors π on $\mathcal{A}$ defined as the collection

$$\Pi := \left\{\pi \in \mathcal{M}(\mathcal{A}) \,\middle|\, \mathbb{E}_{\mu\sim\pi}\left[\mathbb{E}_\mu[X]\right] = q\right\}.$$

and the map $\mathbb{D}^n$ corresponding to the observation of n i.i.d. samples of μ. For $q \in (0, a)$, let $\mathfrak{Q}$ be the set of probability measures $\mathbb{Q}$ on $[0, 1]$ such that $\mathbb{E}_{q' \sim \mathbb{Q}}[q'] = q$. Let $\mathbb{Q}$ be the probability measure on $[0, 1]$ with probability density function $p(x) = (1 - q)/q$ on $[0, q]$ and $p(x) = q/(1 - q)$ on $(q, 1]$. It is easy to check that $\mathbb{Q} \in \mathfrak{Q}$, that

$$\mathbb{E}_{q' \sim \mathbb{Q}} \left[\inf_{\mu \in \mathcal{A} : \mathbb{E}_\mu[X] = q'} \prod_{i=1}^{n} \mu[B_i] \right] = 0, \tag{65}$$

and that, for all $\delta > 0$,

$$\mathbb{P}_{q' \sim \mathbb{Q}} \left[\sup_{\mu \in \mathcal{A} : \mathbb{E}_\mu[X] = q', \prod_{i=1}^{n} \mu[B_i] > 0} \mathbb{E}_\mu[X] > 1 - \delta \right] > 0. \tag{66}$$

It follows from Theorem 4.13 that

$$\mathcal{U}\big(\Psi^{-1}\mathfrak{Q} \big| B\big) = 1. \tag{67}$$

5. Bayesian robustness and consistency

It is appropriate at this point to place the results of Sections 3 and 4 in the more well-established context of two key questions about Bayesian inference, namely its *robustness* with respect to perturbations of the prior (and likelihood and observed data), and its frequentist *consistency*. This discussion will also motivate Section 6, where we show that Bayesian inference can be profoundly non-robust even under arbitrarily small local perturbations in total variation and Prokhorov metrics.

5.1. Bayesian robustness

The robust Bayesian viewpoint appears to have been introduced independently by Box [25] and Huber [58]; see e.g. [15, 16] and Chapter 15 of [60] for surveys of the field. In the robust Bayesian approach, a class Π of priors and a class Λ of likelihoods together produce a class of posteriors by pairwise combination through Bayes' rule. Robust Bayesian methods are a subclass of the methods of *imprecise probability*; the idea that the probability of an event need not be a single real number has a history stretching back to Boole [24] and Keynes [65], with more recent and comprehensive foundations laid out in e.g. [68, 100, 105].

One way of generating such a class Π of priors is via a belief function, as in [104] and Dempster–Shafer theory more generally. The belief function framework encompasses prior probabilities whose values are known only on some finite partition of the probability space, and not the whole σ-algebra; classes of ε-contaminated priors can also be represented in this way, as well as classes of locally perturbed priors. The belief function approach has the useful feature that explicit formulae can be given for the lower and upper posterior probabilities of events [104, Theorem 4.1].

Another typical approach to generating a class Π might be to consider a finite-dimensional parametrized class of models. For example, one could consider,

instead of a single Gaussian prior on $\mathbb{R}$ of specified mean and variance, a two-parameter class of Gaussian priors with a range of means and variances, or a three-parameter class of skew-Gaussian priors. Similarly, one might consider a two-parameter class of beta distributions instead of a uniform prior on a bounded interval.

However, a danger in specifying a finite-dimensional class Π of priors is that one is making very strong statements about the form of the priors, particularly with regard to the tails, that cannot be justified based on often-limited amounts of prior information. For example, if all the priors $\pi \in \Pi$ have thin tails, then the class Π will have a very difficult time modeling events that lie in those tails, even when exposed to data from those regions. This problem is particularly important in applied fields such as catastrophe modeling, insurance, and re-insurance, in which the catastrophic events of interest are by definition high-impact low-probability "Black Swan" events: the difference between an exponentially small and an inverse-polynomially small tail can be vitally important. Also, because members of a finite-dimensional parametric family Π of priors often have similar qualitative properties (such as being mutually absolutely continuous), the apparently broader perspective does not not add much to the asymptotic posterior picture in terms of robust consistency, although it does provide a broader understanding given finitely many samples.

Rather than specifying a finite-dimensional Π, it is epistemologically more reasonable to specify a finite-*codimensional* Π, for example by specifying interval bounds on the expected values of finitely many observed test functions (i.e. generalized moment inequalities); this setting encompasses the finite-partition belief function framework mentioned above. Calculation of optimal prior and posterior bounds on quantities of interest is often an exercise in numerical optimization [20, 82, 90] rather than closed-form formulae.

One consequence of Theorems 4.8 and 4.13 is that the very same Bayesian sensitivity analysis framework that produces the robustness results of classical robust Bayesian inference under finite-dimensional classes of priors also leads to brittleness results under finite-codimensional classes of priors, when the set of all priors is infinite dimensional. As illustrated by (8) and Example 4.10, Theorems 4.8 and 4.13 can also be used to obtain robustness/stability results by adding sufficiently strong constraints (at the expense of learning) on the probability of the data in the model class. As discussed in Subsection 1.4, Example 4.10 suggests that posterior stability and learning are antagonistic properties in Bayesian inference under finite information.

5.2. Motivation for Bayesian inconsistency and model misspecification

To motivate Section 6 and interpret the results of this paper in relation to the issue of convergence of posterior values in Bayesian inference we will now analyse and review questions of Bayesian consistency, inconsistency and model misspecification. There is, of course, a large literature on these topics, and we will not attempt to be exhaustive in providing references; rather, our aims are:

first, to give a short reminder on how Bayesian inference is currently employed in Uncertainty Quantification (UQ); second, to identify issues and popular beliefs about what one actually learns from Bayesian inference, and thereby motivate the results of this paper; and, last, to present sufficient references that the interested reader can find technical justification for the formal manipulations of this subsection.

In this subsection, we are interested in estimating $\Phi(\mu^\dagger)$ where Φ is a known *quantity of interest* function and $\mu^\dagger$ is an unknown (or partially known) probability measure on $\mathcal{X}$. For the purposes of exposition, in this subsection, we assume that $\mathcal{X} = \mathbb{R}^k$. One example of a quantity of interest, when $\mathcal{X} = \mathbb{R}$, is $\Phi(\mu^\dagger) := \mu^\dagger[X \geq a]$ (the probability that the random variable X distributed according to $\mu^\dagger$ exceeds the threshold value a). We also assume that we are given n independent samples $d_1, \ldots, d_n$, each distributed according to $\mu^\dagger$.

We will now present the parametric Bayesian answer to this problem. For the purposes of exposition, in this section, we restrict our attention to parametric Bayesian inference. We first introduce $\{\mu(\cdot\,, \theta)\}_{\theta \in \Theta}$ a family of probability distributions on $\mathcal{X}$ parametrized by $\theta \in \Theta$ (and commonly referred to as the *model class*). For the sake of simplicity here we also assume that $\Theta = \mathbb{R}^\ell$. Let

$$\mathcal{A}_0 := \big\{\mu(\cdot\,, \theta) \,\big|\, \theta \in \Theta\big\}.$$

Note that $\mathcal{A}_0$ is a subset of $\mathcal{M}(\mathcal{X})$ that may or may not contain $\mu^\dagger$. If $\mu^\dagger \notin \mathcal{A}_0$, then the model is said to be *misspecified*; otherwise, the model is said to be *well specified*.

We next introduce $p_0 \in \mathcal{M}(\Theta)$, a probability distribution on Θ (the *prior distribution* on θ). Let π_0 be the push-forward (measure) of p_0 under the map $\theta \mapsto \mu(\cdot\,, \theta)$ (see [23, 22], Sections 3.6, 3.7) and observe that π_0 is a probability distribution on $\mathcal{A}_0$, i.e. $\pi_0 \in \mathcal{M}(\mathcal{A}_0)$, and that π_0 is the distribution of the random measure $\mu(\cdot\,, \theta)$ when θ is distributed according to p_0.

The next step is then to estimate $\Phi(\mu^\dagger)$ via conditioning. Let $p_n \in \mathcal{M}(\Theta)$ be the posterior distribution of θ given the observation of the i.i.d. samples $d_1, \ldots, d_n$, as obtained using Bayes' formula, and let π_n be the push-forward of p_n. The Bayesian estimate of $\Phi(\mu^\dagger)$ is therefore

$$\mathbb{E}_{\mu \sim \pi_n}[\Phi(\mu)]. \tag{68}$$

For the purposes of exposition, we assume that the measures $\mu(\cdot\,, \theta)$ and $\mu^\dagger$ are all absolutely continuous with respect to the Lebesgue measure and write $f(\cdot\,, \theta)$ and $f^\dagger$ for their densities, which we assume to be continuous. Similarly, we assume that the measure p_0 is absolutely continuous with respect to the Lebesgue measure and, abusing notation, write p_0 for both the measure p_0 and its (continuous) density, and similarly for $p_n(\cdot\,)$, the posterior density of θ on Θ given the observation the samples $d_1, \ldots, d_n$. We will now examine the convergence properties of the sequence of posterior densities $p_n(\theta)$ as $n \to \infty$. This analysis being classical (see for instance [79] and references therein), our purpose is not to provide rigorous justifications but rather to familiarize the reader with the mechanisms regarding the convergence of posteriors.

We have

$$p_n(\theta) = \frac{p_0(\theta)\prod_{j=1}^n f(d_j,\theta)}{\int_\Theta p_0(\theta')\prod_{j=1}^n f(d_j,\theta')\,\mathrm{d}\theta'} \equiv \frac{p_0(\theta)\prod_{j=1}^n f(d_j,\theta)}{\mathbb{E}_{p_0}[\prod_{j=1}^n f(d_j,\,\cdot\,)]}$$

which we write as

$$p_n(\theta) = \frac{p_0(\theta)e^{nL_n(\theta)}}{\int_\Theta p_0(\theta')e^{nL_n(\theta')}\,\mathrm{d}\theta'} \equiv \frac{p_0(\theta)e^{nL_n(\theta)}}{\mathbb{E}_{p_0}[e^{nL_n(\,\cdot\,)}]},$$

where

$$L_n(\theta) := \frac{1}{n}\sum_{j=1}^n \log f(d_j,\theta).$$

Recall that $\prod_{j=1}^n f(d_j,\theta)$ is commonly known as the *likelihood* and $L_n(\theta)$ as the (*sample*) *average log-likelihood*.

Consistency and the large-sample limit Now observe that if $\log f(d_j,\theta)$ is integrable then it follows from the Law of Large Numbers that $L_n(\theta)$ converges almost surely, as $n\to\infty$, to the *expected log-likelihood* $L(\theta)$ defined by

$$L(\theta) := \int_{\mathcal{X}} f^\dagger(x)\log\big(f(x,\theta)\big)\,\mathrm{d}x. \tag{69}$$

Assuming that $L(\theta)$ has a unique maximizer $\theta^*\in\Theta$ (corresponding to the asymptotic limit of the *maximum likelihood estimator* (MLE), as the number of data points goes to infinity)and that p_0 is strictly positive in every neighborhood of θ^*, it follows under regularity assumptions on f (or local strict convexity in the neighborhood of θ^*) that $p_n(\theta)$ converges, almost surely, as $n\to\infty$, towards a Dirac mass supported at θ^*. Therefore, assuming Φ to be sufficiently regular, the Bayesian posterior estimate of $\Phi(\mu^\dagger)$, i.e.,

$$\int_\Theta \Phi\big(\mu(\,\cdot\,,\theta)\big)p_n(\theta)\,\mathrm{d}\theta, \tag{70}$$

converges almost surely as $n\to\infty$ to

$$\Phi\big(\mu(\,\cdot\,,\theta^*)\big). \tag{71}$$

Note that

$$L(\theta) = \mathrm{Ent}(f^\dagger) - D_{\mathrm{KL}}\big(f^\dagger\big\|f(\,\cdot\,,\theta)\big),$$

where $\mathrm{Ent}(f^\dagger) := -\int_{\mathcal{X}} f^\dagger(x)\log f^\dagger(x)\,\mathrm{d}x$ is the *entropy* of $f^\dagger$ and D_{KL} denotes the *Kullback–Leibler divergence* defined by

$$D_{\mathrm{KL}}\big(f^\dagger\big\|f(\,\cdot\,,\theta)\big) := \mathbb{E}_{x\sim f^\dagger}\left[\log\frac{f^\dagger(x)}{f(x,\theta)}\right].$$

It follows that θ^* is also the minimizer of $D_{\mathrm{KL}}(f^\dagger\|f(\,\cdot\,,\theta))$ with respect to θ, i.e. the MLE θ^* is characterized by the property that $\mu(\,\cdot\,,\theta^*)$ is the distribution having minimal relative entropy to $\mu^\dagger$ in the model class $\{\mu(\,\cdot\,,\theta)\}_{\theta\in\Theta}$.

An immediate consequence of this observation is the fact if the model is not misspecified, i.e. if $\mu^\dagger$ is an element $\mu(\,\cdot\,,\theta^\dagger)$ of the model class, then $\theta^*=\theta^\dagger$,

$\mu(\cdot\,, \theta^*) = \mu^\dagger$, and the Bayesian estimate (70) is asymptotically exact in the limit as $n \to \infty$. In this situation, the Bayesian estimate is said to be *consistent*.

This convergence result is known as the Bernstein–von Mises Theorem (see for instance [79, Theorem 5]) or as the Bayesian Central Limit Theorem, since the limiting posterior can even be described in a more refined way as being asymptotically normal and not just a point mass. The condition that every open neighborhood of $\theta^\dagger$ has strictly positive p_0-probability (or, even more strongly, that the prior be globally supported) has been named *Cromwell's Rule*[3] by Lindley [72].

Recent results [30, 61, 71, 79] on the Bernstein–von Mises phenomenon show a notable dependence of the validity of the Bernstein–von Mises property upon subtle geometrical and topological details, and regularity properties of the model and the data-generating distribution. Therefore, it is to be expected that any general stability condition for Bayesian inference would have to take account of such factors.

What happens when the model is misspecified? To provide an illustrative answer to this question, consider the family of Gaussian models $\{f(\cdot\,, \theta) \mid \theta = (c, \sigma) \in \mathbb{R} \times \mathbb{R}_+\}$, where

$$f(x, c, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-c)^2}{2\sigma^2}\right).$$

What will happen when this model is exposed to data coming from a potentially non-Gaussian truth $\mu^\dagger$, with density $f^\dagger$, that has a well-defined mean $c^\dagger$ and standard deviation $\sigma^\dagger$? By the above considerations, θ^* maximizes the expected log-likelihood (69) with respect to θ, and the expected log-likelihood is simply

$$L(\theta) = -\int_{\mathbb{R}} f^\dagger(x) \frac{(x-c)^2}{2\sigma^2}\,\mathrm{d}x - (\log\sigma)\int_{\mathbb{R}} f^\dagger(x)\,\mathrm{d}x - \log\sqrt{2\pi}. \tag{72}$$

A quick calculation using partial derivatives shows that $\theta^* = (c^*, \sigma^*)$ maximizes (72) if and only if $c^* = c^\dagger$ and $\sigma^* = \sigma^\dagger$. That is, the Bayesian estimate (68) of $\Phi(\mu^\dagger)$, for *any* distribution $\mu^\dagger$ of mean $c^\dagger$ and standard deviation $\sigma^\dagger$, converges almost surely as the number of sample data goes to infinity, towards $\Phi(\mu(\cdot\,, (c^\dagger, \sigma^\dagger)))$, where $\mu(\cdot\,, (c^\dagger, \sigma^\dagger))$ is the unique Gaussian distribution on $\mathbb{R}$ with mean $c^\dagger$ and standard deviation $\sigma^\dagger$.

However, now there is a problem: there are many different probability distributions μ on $\mathbb{R}$ that have the same first and second moments as $\mu^\dagger$ but have, say, different higher-order moments, or different quantiles. Predictions of those other moments or quantiles using $\mu(\cdot\,, (c^\dagger, \sigma^\dagger))$ can be inaccurate by orders of magnitude. A trivial, albeit extreme, example is furnished by $\Phi(\mu) := \mathbb{E}_\mu[|X - c_\mu| \geq t\sigma_\mu]$ (where c_μ and σ_μ denote the mean and standard deviation of μ). Under

[3] Since the posterior cannot possibly concentrate on a point outside the support of the prior, having a globally-supported prior and hence not ruling out a priori any $\theta \in \Theta$ as a possible $\theta^\dagger$ can be seen as a Bayesian version of Oliver Cromwell's famous injunction to the Synod of the Church of Scotland in 1650: "I beseech you, in the bowels of Christ, think it possible that you may be mistaken."

the Gaussian model, (defining $\operatorname{erf}(z) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{s}}\, dt$ as the *error function*)

$$\mathbb{P}\big[|X - c_\mu| \geq t\sigma_\mu\big] = 1 + \operatorname{erf}\left(-\frac{t}{\sqrt{2}}\right),$$

whereas the extreme cases that prove the sharpness of Chebyshev's inequality — in which the probability measure is a discrete measure with support on at most three points in $\mathbb{R}$ — have

$$\mathbb{P}\big[|X - c_\mu| \geq t\sigma_\mu\big] = \min\left\{1, \frac{1}{t^2}\right\}.$$

In the case of the archetypically rare "6σ event", the ratio between the two is approximately 1.4×10^7. This is, of course, an almost perversely extreme comparison: it would be obvious to any observer with only moderate amounts of sample data that the data were being drawn from a highly non-Gaussian distribution. However, it is not inconceivable that the true distribution $\mu^\dagger$ has a Gaussian-looking bulk but tails that are significantly fatter than those of a Gaussian, and the difference may be difficult to establish using reasonable amounts of sample data; yet, it is those tails that drive the occurrence of "Black Swans", catastrophically high-impact but low-probability outcomes. The results of this paper suggest that this situation is generic, and cannot be avoided no matter how many moments or integrals of arbitrary test functions of the truth $\mu^\dagger$ are matched nor how "close" $\mu^\dagger$ is to the class $\{\mu(\,\cdot\,, \theta)\}_{\theta\in\Theta}$.

5.3. Bayesian inconsistency and model misspecification

To quote [79], "[w]hile for a Bayesian statistician the analysis ends in a certain sense with the posterior, one can ask interesting questions about the the properties of posterior-based inference from a frequentist point of view." Many of these questions are asymptotic in nature: for example, in the limit of infinitely many independent $\mu^\dagger$-distributed samples, will the posterior converge in a suitable sense to $\mu^\dagger$ regardless of the initial choice of prior π? This property is referred to as *consistency*[4]; a general survey of consistency results is found in [99]. As noted above, the consistency theorem is generically known as the Bernstein–von Mises theorem [19, 96], although the earliest rigorous proofs are due to Doob [38] and Le Cam [69].

Unfortunately, Cromwell's Rule is only necessary, and not sufficient, to ensure consistency. In fact, consistency is far from being a generic property, and once the probability space contains infinitely many points (and hence any parameter space Θ that parametrizes all probability measures on that probability space is infinite-dimensional), inconsistency is not the exception, but the rule [36]. In [48, Sec. 5], Freedman considered a countable index set $\mathbb{N} := \{1, 2, \dots\}$ and the

[4]Sometimes the term *frequentist consistency* is used, reflecting the fact that it lies outside the strict Bayesian worldview.

parameter space

$$\Theta := \left\{ \theta \colon \mathbb{N} \to [0,1] \,\middle|\, \sum_{i \in \mathbb{N}} \theta(i) = 1 \right\}.$$

Each θ gives rise to a probability distribution $\mathbb{P}_\theta = \mu(\cdot\,, \theta)$ under which the observations $X_1, X_2, \ldots$ are IID with $\mathbb{P}_\theta[X_n = i] = \theta(i)$. The problem is assumed to be well-specified, so that one particular $\theta^\dagger \in \Theta$ is considered to be the "true" parameter value, and the frequentist data-generating distribution is $\mu^\dagger = \mathbb{P}_{\theta^\dagger} = \mu(\cdot\,, \theta^\dagger)$. Theorem 5 of [48] shows that, when $\mathrm{supp}(\mu^\dagger)$ is infinite, given any "spurious" probability distribution $\mathbb{Q} = \mathbb{P}_q$, there exists a prior probability measure π on Θ that has $\theta^\dagger$ in its support, such that the posterior of π $\mu^\dagger$-a.s. concentrates on q in the limit of observing infinitely many i.i.d. $\mu^\dagger$-distributed samples. In fact, there is a prior that gives positive mass to every open subset of Θ but yields consistent posterior estimates for only a first-category set of possible "true" (data-generating) parameter values $\theta^\dagger$.

There are conditions on priors that can ensure frequentist consistency in infinite-dimensional or non-parametric contexts, e.g. the tail-free priors introduced by Freedman in [48] and hybrid Bayesian–frequentist tools such as Dirichlet process priors [52]. However, while the collection of "bad" priors that lead to inconsistent results is measure-theoretically small [38, 28], it is topologically generic [49].

Remark 5.1. It is probably fair to say that, despite their popularity and documented successes, Bayesian methods have always attracted some degree of controversy and opposition: see e.g. [51] and rejoinders for a recent academic discussion, and [73, 78] for less formal treatments. Often, this opposition is philosophical in nature, particularly with regard to the subjective interpretation of the probabilities involved, which is something that remains counter-intuitive to many commentators: see [44, par. 35 & 37] for a recent example in law. However, there are also analytical reasons to be careful about the application of Bayesian methods [88, 76, 43]. It is, in fact, now well understood that Bayesian methods may fail to converge or may converge towards the wrong solution if the underlying probability mechanism allows an infinite number of possible outcomes [35] and that, in these non-finite-probability-space situations, this lack of convergence (commonly referred to as *Bayesian inconsistency*) is the rule rather than the exception [36]. There is now a wide literature of positive [19, 30, 38, 67, 69, 96, 92] and negative results [12, 35, 48, 47, 61, 71] on the consistency properties of Bayesian inference in parametric and non-parametric settings, and an emerging understanding of the fine topological and geometrical properties that determine (in)consistency.

It is important to appreciate that the requirement of positive prior mass in every neighborhood of the true distribution depends upon the topology placed upon $\mathcal{M}(\mathcal{X})$. For example, Schwartz [86] shows that every π that puts positive mass on all Kullback–Leibler (relative entropy) neighborhoods of $\mu^\dagger$ is weakly consistent. On the other hand, Freedman [48] and Diaconis & Freedman [35] show that π may put positive mass on all weak neighborhoods of $\mu^\dagger$ and still fail

to be weakly consistent — e.g. by not being tail-free. Nor are results limited to *weak* convergence of the posterior to $\mu^\dagger$. For example, [9] shows that consistency holds in the Hellinger distance if π puts positive mass on all Kullback–Leibler neighborhoods of $\mu^\dagger$ and certain smoothness and tail conditions are satisfied; see [98, 101] for further results on Hellinger and Kullback–Leibler consistency. The amount of prior probability mass that lies Kullback–Leibler-close to the truth, quantified using a notion called *thickness*, can be used to quantify the convergence properties of Bayes estimates [1, 2, 74]. However, it is important to note that, in the infinite-dimensional contexts that are increasingly subject to Bayesian analyses, results like the Feldman–Hájek dichotomy [45, 56] suggest that probability measures are 'usually' mutually singular and 'rarely' mutually absolutely continuous, and so the Kullback–Leibler neighborhoods of $\mu^\dagger$ are 'small' sets that are 'unlikely' to intersect the model class.

The situation in which there is no $\theta^\dagger \in \Theta$ such that $\mu^\dagger = \mu(\cdot, \theta^\dagger)$ is referred to as *model misspecification*. The consistency and other asymptotic properties of misspecified models appear to have first been considered by Berk [17, 18] and Huber [59]. See [66, 67] for a recent contribution, and [74] for convergence rates.

> "In practice, Bayesian inference is employed under misspecification *all the time*, particularly so in machine learning applications. While sometimes it works quite well under misspecification [21, 66], there are also cases where it does not [31, 50], so it seems important to determine precise conditions under which misspecification is harmful — even if such an analysis is based on frequentist assumptions." [53]

There is a reasonable popular belief that gross misspecification of the model will be detected by some means before engaging in a serious Bayesian analysis; indeed there do exist tests [57, 106] for model misspecification, but it is important to note that while one *can* determine that the model is misspecified, one *cannot* be sure that the model is well-specified. There is also an understandable popular belief that these tests mean that one need only be concerned with the situation of "mild misspecification", and that provided $\mu^\dagger$ lies "close enough" to the model class $\{\mu(\cdot, \theta)\}_{\theta \in \Theta}$, the posterior estimates will still converge to a usefully informative limit.

Remark 5.2. This belief echoes G. E. P. Box's statement [26, p. 424] that "essentially, all models are wrong, but some are useful" and question [26, p. 74] "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful?"

In terms of the above discussion, one purpose of this paper is to explore the extent to which one can simultaneously have *robust* Bayesian analyses that produce *consistent* answers, given that the models used (both priors and likelihoods) are certain to be *misspecified* to some degree. Can one be "just a little bit wrong" in terms of model misspecification? Our results suggest that the answer is negative within the classical framework of Bayesian Sensitivity analysis, when "closeness" is measured in terms of total variation and Prokhorov metrics or in terms of a finite (but possibly large) number of marginals of the data generating distribution.

In particular, one aim of Section 6 is to show that this belief is wrong if "mild misspecification" is measured using the Prokhorov or the total variation metrics, the number of samples is finite (but possibly arbitrarily large), and if convergence is required to hold uniformly in an arbitrarily small neighborhood of the model.

Remark 5.3. It is known from the Bernstein–von Mises theorem [19, 96] that, in finite-dimensional situations, posterior values converge towards the quantity of interest if the prior distribution has strictly positive mass in every neighborhood of the truth (see also [69, 79]). It is also known that "even for the simplest infinite-dimensional models, the Bernstein–von Mises theorem does not hold" [32, 47]. This possible lack of convergence, referred to as the consistency problem, has been at the center of a debate between frequentists and Bayesians. We quote Diaconis and Freedman [35] (see also [36])

> "If the underlying mechanism allows an infinite number of possible outcomes (e.g., estimation of an unknown probability on the integers), Bayes estimates can be inconsistent: as more and more data comes in, some Bayesian statisticians will become more and more convinced of the wrong answer."

What is the significance of Theorem 4.13 in that discussion? To answer this question, consider Example 4.9 (and 4.19), in which one is interested in estimating the probability (under the unknown measure $\mu^\dagger$) that X exceeds a after observing n independent samples. We already know from [35, 32] that placing priors on the infinite-dimensional space $\mathcal{A} = \mathcal{M}[0,1]$ of probability measures on $[0,1]$ is unlikely to lead to Bayesian posteriors that will converge towards the true value as more and more data comes in. One strategy to circumvent this lack of convergence would be to consider a finite-dimensional subset of $\mathcal{A}$, i.e. a family (μ_λ) of probability measures on $[0,1]$ indexed by a finite-dimensional parameter $\lambda \in \mathbb{R}^k$, put a strictly positive prior p on $\lambda \in \mathbb{R}^k$, and then invoke the Bernstein–von Mises theorem to guarantee the convergence of posterior values.

However, the Bernstein–von Mises theorem requires that the true distribution under which the data is sampled belongs to $\{\mu_\lambda \mid \lambda \in \mathbb{R}^k\}$, the parametrized finite-dimensional subset of $\mathcal{A}$. What happens when this is not the case, i.e. the situation of *misspecification*? Write π_p for the push-forward of the prior p on $\lambda \in \mathbb{R}^k$ to a prior on $\mathcal{A}$ under the map $\lambda \mapsto \mu_\lambda$. Assume that the data have been sampled from $\pi^\dagger \cdot \mathbb{D}$ where $\pi^\dagger$ is the (frequentist) true distribution. Here Theorem 4.13, as illustrated in Example 4.16, can be used to show that the posterior values of the quantity of interest under π_p and $\pi^\dagger$ may lie near the opposite extreme values of Φ in $\mathcal{A}$ even if (1) $\pi^\dagger$ is a Dirac mass on a measure $\mu^\dagger \in \mathcal{A}$; (2) the number of independent samples is large; and (3) k is large and k moments of $\mu^\dagger$ and μ_{λ^*} are equal for some $\lambda^* \in \mathbb{R}^k$.

Remark 5.4. One popular method for detecting failure of convergence under model misspecification is to divide the data into data used for calibrating the parameters of the model and data used for validating the accuracy or predictability of the (calibrated) model. This approach, oftentimes described as "frequentist" [55, 11], could be used to validate Bayesian calculations [43]. Although the de-

tection (of the lack of predictability of the model) is asymptotically robust, it requires the availability of sufficient data.

6. Brittleness under local misspecification

The purpose of this section is to present brittleness results with respect to local perturbations in the total variation and Prokhorov metrics. Thus, whereas the examples given for Theorem 4.13 highlighted that no finite number of common moments would be sufficient to constrain two priors to give nearby posterior value for the quantity of interest, this section shows that closeness in the TV and Prokhorov metrics is also insufficient to ensure robustness.

We now establish a corollary to the proof of Theorem 4.13 which we will then use to establish an extreme brittleness theorem for a model with local misspecification. Recall that, for a map $\Psi\colon \mathcal{A} \to \mathcal{Q}$, a map $\psi\colon \Psi(\mathcal{A}) \to \mathcal{A}$ is called a *section* of Ψ if $\Psi \circ \psi(q) = q$ for all $q \in \Psi(\mathcal{A})$.

Theorem 6.1. *Let $\mathcal{A}$ be a Suslin space, let $\Phi\colon \mathcal{A} \to \mathbb{R}$ be measurable, let $\mathcal{Q}$ be a separable and metrizable space, and let $\Psi\colon \mathcal{A} \to \mathcal{Q}$ measurable. Let $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ be such that* $\operatorname{supp}(\mathbb{Q}) \subseteq \Psi(\mathcal{A})$ *for all $\mathbb{Q} \in \mathfrak{Q}$. Let the data space $\mathcal{D}$ be metrizable and consider $B \in \mathcal{B}(\mathcal{D})$. Assume that $\mathbb{D}$ is such that all the level sets of Ψ go to zero, in the sense that*

$$\inf_{\mu \in \Psi^{-1}(q)} \mathbb{D}(\mu)[B] = 0, \quad \text{for all } q \in \Psi(\mathcal{A}). \tag{73}$$

Then for any positive measurable section ψ of Ψ, positive in the sense that

$$\mathbb{D}(\psi(q))[B] > 0, \quad \text{for all } q \in \Psi(\mathcal{A}), \tag{74}$$

it follows that

$$\mathcal{U}\big(\Psi^{-1}\mathfrak{Q}\big|B\big) \geq \mathfrak{Q}^{\infty}(\Phi \circ \psi). \tag{75}$$

where $\mathfrak{Q}^{\infty}(\Phi \circ \psi)$ is the essential supremum

$$\mathfrak{Q}^{\infty}(\Phi \circ \psi) := \sup_{\mathbb{Q} \in \mathfrak{Q}} \inf \big\{ r \in \mathbb{R} : \mathbb{Q}[\Phi \circ \psi > r] = 0 \big\}. \tag{76}$$

See Figure 3 for an illustration of Theorem 6.1.

We now use Theorem 6.1 to develop a brittleness theorem for a model with local misspecification. To that end, let $\mathcal{X}$ be a Polish space so that, by [4, Thm. 15.15], $\mathcal{M}(\mathcal{X})$ endowed with the weak topology is Polish. Moreover, by [40, Thm. 11.3.3], we know that if we select a complete consistent metric d for $\mathcal{X}$, then the Prokhorov metric $d_{\mathcal{M}}$ defined by

$$d_{\mathcal{M}}(\mu_1, \mu_2) := \inf \{ \varepsilon > 0 \,|\, \mu_1(A) \leq \mu_2(A^{\varepsilon}) + \varepsilon \text{ for all } A \in \mathcal{B}(\mathcal{X}) \},$$

where

$$A^{\varepsilon} := \{ x \in \mathcal{X} \,|\, d(x, x') < \varepsilon \text{ for some } x' \in A \}$$

is the ε neighborhood of A, metrizes the weak topology on $\mathcal{M}(\mathcal{X})$. Moreover, Prokhorov's theorem [40, Cor. 11.5.5] asserts that the Prokhorov metric $d_{\mathcal{M}}$ is a complete metric for the Polish space $\mathcal{M}(\mathcal{X})$. For $\alpha > 0$, $\mu \in \mathcal{M}(\mathcal{X})$, let

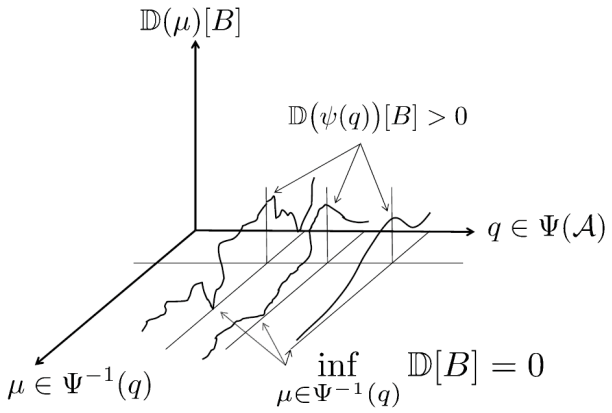


FIG 3. *Illustration of Conditions* (73) *and* (74) *of Theorem 6.1. If, for some data map* $\mathbb{D} \in \mathfrak{D}$*, all level sets of* Ψ *go to zero (i.e. for all* $q \in \Psi(\mathcal{A})$*,* $\inf_{\mu \in \Psi^{-1}(q)} \mathbb{D}(\mu)[B] = 0$*), then, for any positive section* ψ *of* Ψ *(i.e.* $\Psi \circ \psi(q) = q$ *and* $\mathbb{D}(\psi(q))[B] > 0$ *for* $q \in \Psi(\mathcal{A})$*), the least upper bound on posterior values is bounded from below by the essential supremum of* $\Phi \circ \psi$*.*

$B_\alpha(\mu) := \{\mu' \in \mathcal{M}(\mathcal{X}) \mid d_{\mathcal{M}}(\mu, \mu') < \alpha\}$ be the open ball of Prokhorov radius α about μ.

Let Θ be a Polish space and let the model define a map

$$\mathcal{P} \colon \Theta \to \mathcal{M}(\mathcal{X}).$$

As in Section 5.2, the image $\mathcal{P}(\Theta)$ is referred to as the (Bayesian) *model class*.

Remark 6.2. When $\mathcal{P}$ is continuous, it follows from the definition [5, Sec. 3.2] of an analytic set that the the image $\mathcal{P}(\Theta) \subseteq \mathcal{M}(\mathcal{X})$ is analytic, and since the range space $\mathcal{M}(\mathcal{X})$ is Polish it follows that $\mathcal{P}(\Theta)$ is Suslin. Actually, continuity is not required, since [5, Thm. 3.3.4] implies that if $\mathcal{P}$ is measurable, then the image $\mathcal{P}(\Theta)$ is Suslin. If, in addition, $\mathcal{P}$ is injective, then Suslin's Theorem [5, Thm. 3.2.3] implies that $\mathcal{P}(\Theta)$ is Borel.

Assume that $\mathcal{P}$ is measurable and denote its image by $\mathcal{A}_0 := \mathcal{P}(\Theta)$. Let $\pi_\Theta \in \mathcal{M}(\Theta)$ be a prior distribution on Θ and let $\pi_0 := \mathcal{P}\pi_\Theta \in \mathcal{M}(\mathcal{A}_0)$ be its pushforward. Let $\Phi_0 \colon \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ be a measurable quantity of interest. We are interested in estimating Φ_0 using the prior π_0 and our purpose is to show the extreme brittleness of this estimation under arbitrarily small perturbations of the model class $\mathcal{A}_0$ in both the Prokhorov and total variation metrics.

For conditioning on observations, let the data space be $\mathcal{D} := \mathcal{X}^n$, and consider the n-i.i.d. sample data map $\mathbb{D}_0^n \colon \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X}^n)$ defined by

$$\mathbb{D}_0^n \mu := \mu^n, \quad \mu \in \mathcal{M}(\mathcal{X}). \tag{77}$$

For $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$, dropping the notational dependence, denote the rectangle about x^n by

$$B_\delta^n := \prod_{i=1}^n B_\delta(x_i), \tag{78}$$

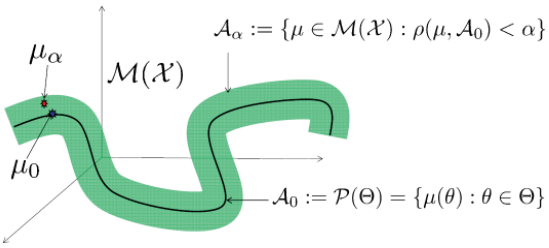


FIG 4. *The original model class $\mathcal{A}_0$ (black curve) is enlarged to its metric neighborhood $\mathcal{A}_\alpha$ (shaded). This procedure determines perturbations $\mu_\alpha \in \mathcal{A}_\alpha$ of the original random measure $\mu_0 \in \mathcal{A}_0$.*

where $B_\delta(x_i)$ is the open ball of radius δ about x_i. Observe that the prior value of Φ_0 under π_0 is $\mathbb{E}_{\pi_0}[\Phi_0]$ and its posterior value under the observation $d \in B_\delta^n$ is $\mathbb{E}_{\pi_0 \odot_{B_\delta^n} \mathbb{D}_0^n}[\Phi_0]$.

To define α-perturbations of the model class $\mathcal{A}_0$ in Prokhorov metric, we introduce, for $\alpha > 0$ the α-neighborhood $\mathcal{A}_\alpha \subseteq \mathcal{M}(\mathcal{X})$ of $\mathcal{A}_0$ defined by (see Figure 4)

$$\mathcal{A}_\alpha := \bigcup_{\mu \in \mathcal{A}_0} B_\alpha(\mu). \tag{79}$$

It is easy to see that the ball fibration (see Remark 6.8)

$$\mathcal{A} := \{(\mu_1, \mu_2) \in \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \,|\, \mu_1 \in \mathcal{A}_0, \mu_2 \in B_\alpha(\mu_1)\} \tag{80}$$

of the set of balls about points of $\mathcal{A}_0$ projects to

$$P_0\mathcal{A} = \mathcal{A}_0 \tag{81}$$

$$P_\alpha\mathcal{A} = \mathcal{A}_\alpha \tag{82}$$

where $P_0 \colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ is the projection onto the first component and P_α the projection onto the second. The naturally induced set of priors corresponding to $\pi_0 \in \mathcal{M}(\mathcal{A}_0)$ is therefore the set $\Pi_\alpha \subset \mathcal{M}(\mathcal{A}_\alpha)$ defined by

$$\Pi_\alpha := \big\{\pi_\alpha \in \mathcal{M}(\mathcal{A}_\alpha) \big| \exists \pi \in \mathcal{M}(\mathcal{A}) \text{ with } P_0\pi = \pi_0 \text{ and } P_\alpha\pi = \pi_\alpha\big\}. \tag{83}$$

Remark 6.3. Observe that each element $\pi_\alpha \in \Pi_\alpha$ is the distribution of a random measure μ_2 on $\mathcal{A}_\alpha$ such that: (i) there exists a random measure $\mu_1 \in \mathcal{A}_0$ with distribution π_0 (that of the model); (ii) (μ_1, μ_2) is jointly measurable; and (iii) with probability one the Prokhorov distance from μ_2 to μ_1 is less than α, i.e. $d_{\mathcal{M}}(\mu_1, \mu_2) < \alpha$. Observe in particular that $\pi_0 \in \Pi_\alpha$.

Our main result is provided in Theorem 6.9 but for the sake of clarity we will first give this result in the following (simpler) form.

Theorem 6.4. *Using the notations introduced above and the data map (77), let Π_α be defined as in (83). If*

$$\lim_{\delta \downarrow 0} \sup_{x \in \mathcal{X}} \sup_{\theta \in \Theta} \mathcal{P}(\theta)[B_\delta(x)] = 0, \tag{84}$$

then, for all $\alpha > 0$ *there exists* $\delta_c(\alpha) > 0$ *such that for all* $0 < \delta < \delta_c(\alpha)$*, all* $n \in \mathbb{N}$*, and all* $(x_1, \ldots, x_n) \in \mathcal{X}^n$

$$\mathcal{U}(\Pi_\alpha | B_\delta^n) \geq \operatorname{ess\,sup}_{\pi_0}(\Phi_0),$$

where

$$\operatorname{ess\,sup}_{\pi_0}(\Phi_0) := \inf\{r > 0 \mid \pi_0[\phi_0 > r] = 0\},$$

and with similar expressions for the lower bounds $\mathcal{L}$*.*

Remark 6.5. Theorem 6.4 implies the extreme brittleness of Bayesian inference under local misspecification. Indeed, assume that the model class $\mathcal{A}_0$ is well specified (i.e. it contains the truth $\mu^\dagger$) and that, therefore, the Bayesian estimator described by π_0 is consistent. One may believe that a model $\mathcal{A}_1$ lying in a 'small enough' neighborhood of $\mathcal{A}_0$ should have good convergence properties, Theorem 6.4 and Remark 6.3 invalidate this belief, at least as far as the TV and Prokhorov notions of 'small enough' are concerned. Using the notations of Remark 6.3, observe in particular that an unscrupulous practitioner may design a model corresponding to a random measure μ_2 such that the distance between μ_1 (the well specified model) and μ_2 is a.s. at most α (where α is arbitrarily small) and the posterior value using the random measure μ_2 is as distant as possible from the posterior value using μ_1 irrespective of the sample size n.

Remark 6.6. Observe that the condition (84) is extremely weak and satisfied for most Bayesian models. This condition can in fact be made weaker by replacing it with the assumption that for n sufficiently large it holds true that for all θ, $\mathcal{P}(\theta)$ does not contain a Dirac mass in each ball $B_\delta(x_i)$ (i.e. on the sample data when $\delta \downarrow 0$). We also note that the proof of Theorem 6.4 does not require the samples to be i.i.d., in particular, the same results can be obtained with coupled samples, if, for instance, the data map $\mathbb{D}_0^n$ is replaced by a data map $\mathbb{D}$ such that $C_1^n \prod_{i=1}^n \mu(A_i) \leq \mathbb{D}(\mu)[A_1 \times \cdots \times A_n] \leq C_2^n \prod_{i=1}^n \mu(A_i)$ for strictly positive constants C_1 and C_2.

Remark 6.7. Theorem 6.4 is a corollary of Theorem 6.9 and the proof of Theorem 6.9 shows that, if Θ is compact and $\mathcal{P}$ is continuous and $\Phi(\mu) := \mu(A)$ for some fixed $A \in \mathcal{B}(\mathcal{X})$, then the result of Theorem 6.4 also holds when using the total variation distance d_{TV} instead of the Prokhorov distance, which produces a much smaller neighborhood.

However, in this metric $\mathcal{M}(\mathcal{X})$ in general is not separable and this introduces measurability difficulties. These difficulties can be overcome somewhat when Θ is compact and $\mathcal{P}$ is continuous, since the image of a compact set under a continuous map is compact and therefore measurable. Moreover, validation or certification type quantities of interest defined by $\Phi(\mu) := \mu(A)$ for some fixed $A \in \mathcal{B}(\mathcal{X})$ are easily seen to be continuous and therefore measurable. Moreover, because of continuity,

$$\Pi_0^\infty(\Phi_0) \approx \Pi_\alpha^\infty(\Phi_0).$$

Our motivation in working mainly with the Prokhorov metric lies in the fact that we also seek to lay down measurability foundations for the scientific computation

of optimal statistical estimators where the unknown quantities are products of functions and measures and for such spaces the total variation metric is too strong for the measurability of standard quantities of interest.

We will now give a more general version of Theorem 6.4 and elaborate on the objects entering in its formulation. We start with $\Pi_\Theta \subseteq \mathcal{M}(\Theta)$, a set of admissible priors and let

$$\Pi_0 := \mathcal{P}\Pi_\Theta \subseteq \mathcal{M}(\mathcal{A}_0)$$

denote the push-forward by the model $\mathcal{P}$. We consider the pull-back $\Phi_\Theta := \Phi_0 \circ \mathcal{P}$, of the measurable quantity of interest $\Phi_0 \colon \mathcal{M}(\mathcal{X}) \to \mathbb{R}$, to a measurable quantity of interest $\Phi_\Theta \colon \Theta \to \mathbb{R}$. Then the change of variables formula [40, Thm. 4.1.11] implies that, for $\pi_\Theta \in \mathcal{M}(\Theta)$,

$$\mathbb{E}_{\pi_\Theta}[\Phi_\Theta] = \mathbb{E}_{\pi_\Theta}[\Phi_0 \circ \mathcal{P}] = \mathbb{E}_{\mathcal{P}\pi_\Theta}[\Phi_0]$$

whenever either side is well defined. Therefore, taking suprema and infima, we obtain

$$\begin{aligned} \mathcal{U}(\Pi_\Theta) &= \mathcal{U}(\Pi_0), \\ \mathcal{L}(\Pi_\Theta) &= \mathcal{L}(\Pi_0), \end{aligned}$$

where we note that the quantity of interest implicit in these definitions is determined by the argument. For $\alpha > 0$, define $\mathcal{A}_\alpha$, $\mathcal{A}$, P_0 and P_α as in (79), (80), (81) and (82).

Remark 6.8. Using the affine convexity of $\mathcal{M}(\mathcal{X})$, one can show that $\mathcal{A}$ is indeed a Hurewicz fibration, in that it has the homotopy lifting property, see e.g. [91, p. 66]. Let

$$d_{\mathcal{M}}^{-1}(< \alpha) := \{(\mu_1, \mu_2) \mid d_{\mathcal{M}}(\mu_1, \mu_2) < \alpha\}$$

denote the set of all pairs of measures at Prokhorov distance at most α from one another. Since $d_{\mathcal{M}} \colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ is continuous, it follows that $d_{\mathcal{M}}^{-1}(< \alpha)$ is open and therefore Borel. In addition, since $\mathcal{A}_0 \subseteq \mathcal{M}(\mathcal{X})$ is Suslin it follows that $\mathcal{A}_0 \times \mathcal{M}(\mathcal{X}) \subseteq \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ is Suslin. Therefore, since $\mathcal{A} = d_{\mathcal{M}}^{-1}(< \alpha) \cap (\mathcal{A}_0 \times \mathcal{M}(\mathcal{X}))$, it follows that $\mathcal{A}$ is Suslin.

Observe that the measurable quantity of interest $\Phi_0 \colon \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ acting on the second component of $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$, naturally pulls back to the quantity of interest $\Phi \colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ by $\Phi := \Phi_0 \circ P_\alpha$, and we have $\sup_{\mathcal{A}_\alpha} \Phi_0 = \sup_{\mathcal{A}} \Phi$ and $\inf_{\mathcal{A}_\alpha} \Phi_0 = \inf_{\mathcal{A}} \Phi$, i.e.

$$\begin{aligned} \mathcal{U}(\mathcal{A}_\alpha) &= \mathcal{U}(\mathcal{A}), \\ \mathcal{L}(\mathcal{A}_\alpha) &= \mathcal{L}(\mathcal{A}). \end{aligned}$$

For a subset $\Pi_0 \subseteq \mathcal{M}(\mathcal{A}_0)$, the projection identity (81) implies that the set $\Pi := P_0^{-1}\Pi_0$ defined by $P_0^{-1}\Pi_0 := \{\pi \in \mathcal{M}(\mathcal{A}) \mid P_0\pi \in \Pi_0\}$ is the induced

set of probability measures on $\mathcal{A}$. Moreover, for $\pi \in \Pi$, the change of variables formula

$$\mathbb{E}_\pi[\Phi] = \mathbb{E}_\pi[\Phi_0 \circ P_\alpha] = \mathbb{E}_{P_\alpha \pi}[\Phi_0]$$

implies that

$$\begin{aligned}\sup_{\pi\in\Pi} \mathbb{E}_\pi[\Phi] &= \sup_{\pi_\alpha \in P_\alpha \Pi} \mathbb{E}_{\pi_\alpha}[\Phi_0],\\ \inf_{\pi\in\Pi} \mathbb{E}_\pi[\Phi] &= \inf_{\pi_\alpha \in P_\alpha \Pi} \mathbb{E}_{\pi}[\Phi_0],\end{aligned}$$

so that

$$P_\alpha \Pi = P_\alpha P_0^{-1} \Pi_0 \subseteq \mathcal{M}(\mathcal{A}_\alpha)$$

is the induced set of probability measures on $\mathcal{A}_\alpha$. Let us denote this induced set by

$$\Pi_\alpha := P_\alpha P_0^{-1} \Pi_0, \tag{85}$$

so that these equalities become

$$\begin{aligned}\mathcal{U}(\Pi) &= \mathcal{U}(\Pi_\alpha),\\ \mathcal{L}(\Pi) &= \mathcal{L}(\Pi_\alpha).\end{aligned}$$

For conditioning on observations, define $\mathbb{D}_0^n$ as in (77) and pull it back to the data map $\mathbb{D}^n \colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X}^n)$ defined by $\mathbb{D}^n := \mathbb{D}_0^n \circ P_\alpha$. Define B_δ^n as in (78) and recall the definition (43)

$$\mathbb{E}_{\pi \odot \mathbb{D}^n}\left[\Phi \middle| B_\delta^n\right] = \frac{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\left[\Phi(\mu_1,\mu_2)\mathbb{D}^n(\mu_1,\mu_2)[B_\delta^n]\right]}{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\left[\mathbb{D}^n(\mu_1,\mu_2)[B_\delta^n]\right]}.$$

of the conditional expectation and the corresponding (47) upper value

$$\mathcal{U}(\Pi|B_\delta^n) := \sup_{\pi\odot\mathbb{D}^n \in \Pi \odot_{B_\delta^n} \mathbb{D}^n} \mathbb{E}_{\pi\odot\mathbb{D}^n}\left[\Phi\middle|B_\delta^n\right]$$

in terms of the admissible set (45)

$$\Pi_{B_\delta^n} := \left\{\pi \in \Pi : (\pi \cdot \mathbb{D}^n)[B_\delta^n] > 0\right\}$$

of product measures, where the marginal is defined by

$$(\pi \cdot \mathbb{D}^n)[B_\delta^n] := \mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\left[\mathbb{D}^n(\mu_1,\mu_2)[B_\delta^n]\right].$$

Let us indicate the dependence on some measure $\underline{\pi}$ of the essential supremum of some quantity of interest $\underline{\Phi}$ by

$$\underline{\pi}^\infty(\underline{\Phi}) := \inf\{r \in \mathbb{R} \,|\, \underline{\pi}\{\underline{\Phi} > r\} = 0\}$$

and, for a set $\underline{\Pi}$ of measures, let

$$\underline{\Pi}^\infty(\underline{\Phi}) := \sup_{\underline{\pi}\in\underline{\Pi}} \underline{\pi}^\infty(\underline{\Phi}). \tag{86}$$

For $\pi_\alpha = P_\alpha \pi$ with $\pi \in \Pi$, we have

$$\begin{aligned}\pi_\alpha[\Phi_0 > r] &= \big(P_\alpha \pi\big)[\Phi_0 > r] \\ &= \pi[\Phi_0 \circ P_\alpha > r] \\ &= \pi[\Phi > r]\end{aligned}$$

so that we conclude that

$$\Pi^\infty(\Phi) = \Pi^\infty_\alpha(\Phi_0).$$

Let us now quantify a type of regularity for the model $\mathcal{P}$. For $x \in \mathcal{X}$, let $B_0(x) := \{x\}$ and define

$$\mathcal{P}_\infty(\delta) := \sup_{x\in\mathcal{X}} \sup_{\theta\in\Theta} \mathcal{P}(\theta)[B_\delta(x)], \quad \text{for } \delta \geq 0.$$

It is clear that $\mathcal{P}_\infty \colon \mathbb{R}^+ \to [0,1]$ is an increasing function. Moreover, for most parametric families, it is easy to show that $\mathcal{P}_\infty$ is continuous and $\mathcal{P}_\infty(0) = 0$, and for many of them not difficult to find useful upper bounds.

Finally, let us assume that the model $\mathcal{P}$ is positive, in that $\mu(B_\delta(x)) > 0$ for all $\mu \in \mathcal{A}_0$, $x \in \mathcal{X}$, and $\delta > 0$. Theorem 6.4 is a direct consequence of the following theorem.

Theorem 6.9 (Brittleness under Local Misspecification)**.** *With the notation and assumptions above, let* Π_α *be defined as in* (85)*, and let* $\delta > 0$ *and* $0 < \alpha < 1$ *satisfy*

$$\mathcal{P}_\infty(\delta) < \alpha.$$

Then, using $\mathbb{D}_0^n$ *for the distribution of the data, for all integers* $n \geq 1$*,*

$$\mathcal{U}(\Pi_\alpha | B_\delta^n) \geq \Pi_0^\infty(\Phi_0)$$

with similar expressions for the lower bounds $\mathcal{L}$*.*

Remark 6.10. When Cromwell's rule (see Section 5.2) is implemented (i.e. if the prior measure of every non-empty neighborhood is strictly positive), it follows that $\Pi_0^\infty(\Phi_0) = \mathcal{U}(\mathcal{A}_0)$ so that the conclusion of Theorem 6.9 becomes

$$\mathcal{U}(\Pi_\alpha | B_\delta^n) \geq \mathcal{U}(\mathcal{A}_0).$$

Remark 6.11. Theorem 6.9 provides conditions sufficient to guarantee how bad things can get regardless of how many samples are taken. One might hope that when these conditions are not satisfied, that more samples may prove beneficial. However, when the condition

$$\inf_{(\mu,\mu')\in\Psi^{-1}\mu} \mathbb{D}^n(\mu,\mu')[B_\delta^n] = 0, \quad \mu \in \mathcal{A}_0$$

of Theorem 6.1 is only approximately satisfied, the inequality

$$\mathbb{D}^n(\mu,\mu')[B_\delta^n] = \big(\mu'\big)^n[B_\delta^n] = \prod_{i=1}^n \mu'[B_\delta(x_i)]$$

and the quantitative version of Theorem 4.13 (given in [80, Thm. 3.1], see also [80, Rmk. 3.2]) imply that things actually get 'worse' with more samples.

7. Conclusions and further developments

In this paper, we have looked at the robustness of Bayesian Inference in the classical framework of Bayesian Sensitivity Analysis. In that (classical) framework, the data is fixed, and one computes optimal bounds on (i.e. the sensitivity of) posterior values with respect to variations of the prior in a given class of priors. Although robustness is already well established when the class of priors is finite dimensional, we observe that, under general conditions, when the class of priors is finite codimensional, the optimal bounds on posterior values are as large as possible, no matter the number of data points. Our motivation for specifying a finite codimensional class of priors is to look at what classical Bayesian sensitivity analysis would conclude under finite information, and the best way to understand this notion of "brittleness under finite information" is through the simple example provided in Subsection 1.2.

The mechanism causing this "brittleness" has its origin in the fact that, in classical Bayesian sensitivity analysis, optimal bounds on posterior values are computed after the observation of the specific value of the data, and that the probability of observing the data under some feasible prior may be arbitrarily small (the example given in Subsection 1.3 provides an illustration of this phenomenon). This data dependence of *worst priors* is inherent to this classical framework and the resulting brittleness under finite information can be seen as an extreme occurrence of the dilation phenomenon (the fact that optimal bounds on prior values may become less precise after conditioning) observed in classical robust Bayesian inference [103]. Although these worst priors do depend on the data, "look nasty", and make the probability of observing the data very small, they are not "isolated pathologies" but directions of instability (of Bayesian conditioning) and their number increase with the number of data points. The example given in Subsection 1.4 provides an illustration of this point and also suggests that learning and robustness are, to some degree, antagonistic properties: a strong constraint on the probability of the data makes the method robust but learning impossible and, as the constraint is relaxed, learning becomes possible but posterior values become brittle.

Since "brittleness under finite information" appears to be inherent to classical Bayesian Sensitivity Analysis (in which worst priors are computed given the specific value of the data), one may ask whether robustness could be established under finite information by exiting the strict framework of Robust Bayesian Inference and computing the sensitivity of posterior conclusions independently of the specific value of the data. To investigate this question, Hampel and Cuevas' notion of *qualitative robustness* has been generalized in [81] to Bayesian inference based on the quantification of the *sensitivity of the distribution of the posterior distribution* with respect to perturbations of the prior and the data generating distribution, in the limit when the number of data points grows towards infinity. Note that, contrary to classical Bayesian Sensitivity Analysis considered here, in the qualitative formulation the data is not fixed and posterior values are therefore analyzed as dynamical systems randomized through the distribution of the data. To express finite information, the total variation, Prokhorov, and Ky Fan metrics have been used to quantify perturbations and sensitivities.

Since this notion of *qualitative robustness* is established in the limit when the number of data points grows towards infinity, it is natural to expect that the notion of *consistency* (i.e. the property that posterior distributions convergence towards the data generating distribution) will play an important role. Although consistency is primarily a frequentist notion, it is also equivalent to *intersubjective agreement* which means that two Bayesians will ultimately have very close predictive distributions. Therefore, it also has importance for Bayesians. Fortunately, not only are there mild conditions which guarantee consistency, but the Bernstein–von Mises theorem goes further in providing mild conditions under which the posterior is asymptotically normal. The most famous of these are Doob [38], Le Cam and Schwartz [70], and Schwartz [86, Thm. 6.1]. Moreover, the assumptions needed for this consistency are so mild that one can be lead to the conclusion that the prior does not really matter once there is enough data. For example, we quote Edwards, Lindeman and Savage [42]:

> "Frequently, the data so completely control your posterior opinion that there is no practical need to attend to the details of your prior opinion."

To some, the consistency results appeared to generate more confidence than possibly they should. We quote A. W. F. Edwards [41, p. 60]:

> "It is sometimes said, in defence of the Bayesian concept, that the choice of prior distribution is unimportant in practice, because it hardly influences the posterior distribution at all when there are moderate amounts of data. The less said about this 'defence' the better."

[81] shows that the *Edwards defence* is essentially what produces non *qualitative robustness* in Bayesian inference. In particular, the assumptions required for consistency (e.g. the assumption that the prior has Kullback–Leibler support at the parameter value generating the data) are such that arbitrarily small local perturbations of the prior distribution (near the data generating distribution) results in consistency or non-consistency, and therefore have large impacts on the asymptotic behavior of posterior distributions. These mechanisms are different and complementary to those discovered by Hampel and developed by Cuevas, and they suggest that consistency and robustness are, to some degree, antagonistic requirements (a careful selection of the prior is important if both properties, or their approximations, are to be achieved) and also indicate that misspecification generates non *qualitative robustness*.

In conclusion, the exploration of Bayesian inference in a continuous world has revealed both positive and negative results. However, positive results regarding the classical or qualitative robustness of Bayesian inference under finite information have yet to be obtained. To that end, observe that the example provided in Subsection 1.4 suggests that there may be a missing stability condition for Bayesian inference in a continuous world under finite information akin to the CFL condition for the stability of a discrete numerical scheme used to approximate a continuous PDE. Although numerical schemes that do not satisfy the CFL condition may look grossly inadequate, the existence of such perverse examples certainly does not imply the dismissal of the necessity of a stability condition. Similarly, although one may, as in the example provided in Subsec-

tion 1.3, exhibit grossly perverse worst priors, the existence of such priors does not invalidate the need for a study of stability conditions for using Bayesian Inference under finite information. The example of Subsection 1.4 suggests that, in the framework of Bayesian Sensitivity Analysis, under finite information, such a stability condition would strongly depend on how well the probability of the data is known or constrained in the model class in addition to the class of priors and the resolution of the measurements. It is natural to expect that such robustness and stability questions will increase in importance as Bayesian methods increase in popularity due to the availability of computational methodologies and environments to compute the posteriors. Indeed, when posterior distributions are approximated using such methods, the robustness analysis naturally includes not only quantifying sensitivities with respect to the data generating distribution and the choice of prior, but also the analysis of convergence and stability of the computational method. This is particularly true in Bayesian updating where Bayes' rule is applied iteratively and computed posterior distributions become prior distributions for the next iteration. Oftentimes these posterior distributions (which are then treated as prior distributions) are only approximated (e.g. via MCMC methods) and the Brittleness results discussed here and in [81] suggest that having strong convergence (of these MCMC methods) in TV would not be enough to ensure stability. At a higher level, these results appear to suggest that robust inference (in a continuous world under finite information) should be done with reduced/coarse models rather than highly sophisticated/complex models (and the level of "coarseness/reduction" would depend on the available "finite information").

8. Proofs

8.1. Proof of Theorem 3.6

For $q \in \mathbb{R}^n$, define

$$\Pi(q) := \Psi^{-1}\mathfrak{Q} = \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_\pi[\Psi] = q\}$$

and let $\Pi(q,n) := \Pi(q) \cap \Delta(n) \subseteq \Pi(q)$ be the subset consisting of $(n+1)$-fold convex combinations of Dirac masses. Using a 'layercake' approach, we use the fact that

$$\Pi(Z) = \bigcup_{q \in Z} \Pi(q) \quad \text{and} \quad \Pi(Z,n) = \bigcup_{q \in Z} \Pi(q,n),$$

while applying Theorem 3.4 with equality constraints $\Pi(q), q \in \mathbb{R}^n$, and the fact that the supremum over a union is a supremum of suprema to obtain a reduction as follows:

$$\begin{aligned} \mathcal{U}\big(\Pi(Z)\big) &= \mathcal{U}\left(\bigcup_{q \in Z} \Pi(q)\right) \\ &= \sup_{q \in Z} \mathcal{U}\big(\Pi(q)\big) \end{aligned}$$

$$
\begin{aligned}
&= \sup_{q\in Z} \mathcal{U}\big(\Pi(q,n)\big) \\
&= \mathcal{U}\left(\bigcup_{q\in Z} \Pi(q,n)\right) \\
&= \mathcal{U}\big(\Pi(Z,n)\big),
\end{aligned}
$$

which completes the proof. □

8.2. Proof of Lemma 3.10

Since $T \subset \mathcal{Q}$ is a subset of a separable metrizable space, [4, Cor. 3.5] implies that it is itself separable and metrizable. Consider the set-valued map with non-empty values $\Psi^{-1}\colon T \twoheadrightarrow \mathcal{A}$ with graph G defined by

$$G := \{(q,\mu) \in T \times \mathcal{A} \,|\, \Psi(\mu) = q\}\,. \tag{87}$$

Let d be a metric that generated the topology of T and define $h\colon T \times \mathcal{A} \to \mathbb{R}$ by $h(q,\mu) := d(\Psi(\mu), q)$. Then, since d is continuous in each of its arguments, it follows that h is a Carathéodory function, as defined in Definition A.2. Since T is separable and metrizable, Lemma A.3 implies that h is $\mathcal{B}(T)\otimes\mathcal{B}(\mathcal{A})$-measurable. Rewriting Equation (87) as

$$G := \{(q,\mu) \in T \times \mathcal{A} \,|\, h(q,\mu) = 0\}$$

yields that G belongs to $\mathcal{B}(T) \otimes \mathcal{B}(\mathcal{A})$. Lemma A.1 (through the identification $S = \mathcal{A}$, $s = \mu$, $\varphi(t,s) = \Phi(\mu)$) implies that the function $\mathcal{U}\circ\Psi^{-1}\colon T \to \mathbb{R}$ defined for $q \in T$ by $q \mapsto \sup_{\mu\in\Psi^{-1}(q)} \Phi(\mu)$ is $\widehat{\mathcal{B}}(T)$-measurable, thereby establishing the first assertion. The second assertion then follows from the second part of Lemma A.1. □

8.3. Proof of Theorem 3.11

For the first assertion, consider $\mathbb{Q} \in \mathfrak{Q}$. Then, by the second assertion of Lemma 3.10, there exists a $\widehat{\mathcal{B}}(\operatorname{supp}\mathbb{Q})$-measurable section ψ of Ψ, i.e. there exists a $\widehat{\mathcal{B}}(\operatorname{supp}(\mathbb{Q}))$-measurable function $\psi\colon \operatorname{supp}(\mathbb{Q}) \to \mathcal{A}$ such that $\Psi(\psi(q)) = q$ for all $q \in \operatorname{supp}(\mathbb{Q})$. Let $\mathbb{Q}$ also denote its restriction to its support and $\widehat{\mathbb{Q}}$ its completion. Let $\pi := \psi\widehat{\mathbb{Q}} \in \mathcal{M}(\mathcal{A})$, so that, for all $A \in \mathcal{B}(\operatorname{supp}(\mathbb{Q}))$,

$$
\begin{aligned}
\big(\Psi\pi\big)(A) &= \big(\Psi\psi\widehat{\mathbb{Q}}\big)(A) \\
&= \big((\Psi\circ\psi)\widehat{\mathbb{Q}}\big)(A) \\
&= \widehat{\mathbb{Q}}(A) \\
&= \mathbb{Q}(A).
\end{aligned}
$$

Hence, $\Psi\pi = \mathbb{Q}$, establishing the first assertion.

For the main assertion, observe that, for all $\mu \in \mathcal{A}$,

$$\big(\mathcal{U} \circ \Psi^{-1} \circ \Psi\big)(\mu) = \sup_{\mu' : \Psi(\mu') = \Psi(\mu)} \Phi(\mu') \geq \Phi(\mu). \tag{88}$$

Consequently, for $\mathbb{Q} \in \mathfrak{Q}$, the first assertion shows that there is a π such that $\Psi\pi = \mathbb{Q}$, so that a change of variables (Proposition A.7) and the monotonicity properties (Proposition A.4) of these integrals, together with the inequality (88), imply that

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] &= \mathbb{E}_{\Psi\pi}[\mathcal{U} \circ \Psi^{-1}] \\ &= \mathbb{E}_{\pi}[\mathcal{U} \circ \Psi^{-1} \circ \Psi] \\ &\geq \mathbb{E}_{\pi}[\Phi], \end{aligned}$$

and therefore

$$\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] \geq \sup_{\pi \in \Psi^{-1}\mathbb{Q}} \mathbb{E}_{\pi}[\Phi].$$

Consequently,

$$\sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] \geq \sup_{\pi \in \Psi^{-1}\mathfrak{Q}} \mathbb{E}_{\pi}[\Phi] = \mathcal{U}(\Psi^{-1}\mathfrak{Q})$$

and, in particular,

$$\sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] \geq \mathcal{U}(\Psi^{-1}\mathfrak{Q}). \tag{89}$$

To obtain the reverse inequality, for $\delta > 0$ consider $\mathbb{Q} \in \mathfrak{Q}$ and apply Lemma 3.10 to conclude that there exists a δ-optimal $\widehat{\mathcal{B}}(\mathrm{supp}(\mathbb{Q}))$-measurable section of Ψ; that is, a $\widehat{\mathcal{B}}(\mathrm{supp}(\mathbb{Q}))$-measurable function $\psi\colon \mathrm{supp}(\mathbb{Q}) \to \mathcal{A}$ such that $\Psi(\psi(q)) = q$ for all $q \in \mathrm{supp}(\mathbb{Q})$ and $(\Phi \circ \psi)(q) > (\mathcal{U} \circ \Psi^{-1})(q) - \delta$ for all $q \in \mathrm{supp}(\mathbb{Q})$. Now let $\pi := \psi\widehat{\mathbb{Q}} \in \mathcal{M}(\mathcal{A})$, and observe from the proof of the first assertion that $\Psi\pi = \mathbb{Q}$, and therefore $\pi \in \Psi^{-1}\mathbb{Q}$. Therefore, by a change of variables,

$$\begin{aligned} \mathbb{E}_{\pi}[\Phi] &= \mathbb{E}_{\psi\widehat{\mathbb{Q}}}[\Phi] \\ &= \mathbb{E}_{\widehat{\mathbb{Q}}}[\Phi \circ \psi] \\ &> \mathbb{E}_{\widehat{\mathbb{Q}}}[\mathcal{U} \circ \Psi^{-1}] - \delta. \end{aligned}$$

Since, by definition, $\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] := \mathbb{E}_{\widehat{\mathbb{Q}}}[\mathcal{U} \circ \Psi^{-1}]$, it follows that

$$\begin{aligned} \mathcal{U}(\Psi^{-1}\mathfrak{Q}) &= \sup_{\pi \in \Psi^{-1}\mathfrak{Q}} \mathbb{E}_{\pi}[\Phi] \\ &\geq \sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] - \delta. \end{aligned}$$

Since $\delta > 0$ was arbitrary, it follows that

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) \geq \sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}].$$

Recalling the reverse inequality (89), we obtain the main assertion.

The assertion of measure affinity follows from Lemma A.9.

For the assertion (52), define

$$\Pi := \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_\pi[\psi_i] = 0 \text{ for } i = 1, \ldots, n\}\,.$$

Let $\epsilon > 0$. Assume that $\sup_{\pi_+ \in \Pi_+} \mathbb{E}_{\pi_+}[\Phi] > \lambda$ and that $\pi_+ \in \Pi_+$ is such that $\mathbb{E}_{\pi_+}[\Phi] > \lambda$. Observe that $\pi := \pi_+/\pi_+(\mathcal{A})$ is an element of Π that satisfies $\mathbb{E}_\pi[\Phi - \lambda\psi_0] > 0$. Define Π_n as in (27) and apply [82, Thm. 4.1] to $\sup_{\pi \in \Pi} \mathbb{E}_\pi[\Phi - \lambda\psi_0]$ to conclude that there exists $\pi^* \in \Pi_n$ such that $\mathbb{E}_{\pi^*}[\Phi - \lambda\psi_0] > 0$. Since $\Phi - \lambda\psi_0 = (\varphi - \lambda)\psi_0$ and ψ_0 is positive, it also follows that $\mathbb{E}_{\pi^*}[\psi_0] > 0$. Writing $\pi^*_+ := \pi^*/\mathbb{E}_{\pi^*}[\psi_0]$ we obtain that $\pi^*_+ \in \Pi_{+,n}$ and $\mathbb{E}_{\pi^*_+}[\Phi] > \lambda$, which concludes the proof of (52). □

8.4. Proof of Lemma 4.1

Consider the set

$$\mathcal{Y}' := \bigcup \{\mathcal{O}_y \,|\, \mathcal{O}_y \subseteq \mathcal{Y} \text{ is open and } \nu(\mathcal{O}_y) = 0\}\,.$$

First let us show that $E = \mathcal{Y}'$. To see this, first observe that trivially we have $E \subseteq \mathcal{Y}'$. Now suppose that $y \in \mathcal{Y}'$. Then there exists a $y' \in \mathcal{Y}$ and an open $\mathcal{O}_{y'} \ni y'$ such that $y \in \mathcal{O}_{y'}$ and $\nu(\mathcal{O}_{y'}) = 0$. Therefore, $y \in E$ and hence $E = \mathcal{Y}'$.

Now, since $\mathcal{Y}'$ is a union of open sets, it is open and therefore measurable. Moreover, since $\mathcal{Y}$ is strongly Lindelöf, it follows that $\mathcal{Y}'$ is Lindelöf and that the open cover of $\mathcal{Y}'$ by ν-null open sets used in the definition of $\mathcal{Y}'$ has a countable subcover, so that

$$\mathcal{Y}' = \bigcup_{i \in \mathbb{N}} \mathcal{O}_{y_i}$$

where each $\mathcal{O}_{y_i}$ is open and has $\nu(\mathcal{O}_{y_i}) = 0$. It follows that

$$\nu(E) = \nu(\mathcal{Y}') \leq \sum_{i \in \mathbb{N}} \nu(\mathcal{O}_{y_i}) = 0$$

and the proof is finished. □

8.5. Proof of Theorem 4.7

The first assertion, (51), follows by layering the set of positive measures of finite total mass as $\bigcup_{r \in \mathbb{R}_+} \{r\mathcal{M}(\mathcal{A})\}$, using the fact that the supremum over a union is a supremum of suprema, and applying the reduction theorem [82, Thm. 4.1] in $r\mathcal{M}(\mathcal{A})$ separately.

For the second assertion, (52), define

$$\Pi := \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_\pi[\psi_i] = 0 \text{ for } i = 1, \ldots, n\}$$

Let $\epsilon > 0$. Assume that $\sup_{\pi_+ \in \Pi_+} \mathbb{E}_{\pi_+}[\Phi] > \lambda$ and that $\pi_+ \in \Pi_+$ is such that $\mathbb{E}_{\pi_+}[\Phi] > \lambda$. Observe that $\pi := \pi_+/\pi_+(\mathcal{A})$ is an element of Π that satisfies $\mathbb{E}_\pi[\Phi - \lambda\psi_0] > 0$. Defining Π_n as in (27) and applying [82, Thm. 4.1] to

$\sup_{\pi\in\Pi}\mathbb{E}_\pi[\Phi-\lambda\psi_0]$, we deduce that there exists $\pi^*\in\Pi_n$ such that $\mathbb{E}_{\pi^*}[\Phi-\lambda\psi_0]>0$. Since $\Phi-\lambda\psi_0=(\varphi-\lambda)\psi_0$ and ψ_0 is positive, it also follows that $\mathbb{E}_{\pi^*}[\psi_0]>0$. Let $\pi^*_+:=\pi^*/\mathbb{E}_{\pi^*}[\psi_0]$ to obtain that $\pi^*_+\in\Pi_{+,n}$ and $\mathbb{E}_{\pi^*_+}[\Phi]>\lambda$, which concludes the proof of (52). □

8.6. Proof of Theorem 4.8

First, we prove that

$$\mathcal{U}(\Pi(q)|B)=\sup_{\pi_+\in\Pi_+(q)}\mathbb{E}_{\mu\sim\pi_+}\big[\Phi(\mu)\mathbb{D}(\mu)[B]\big],\tag{90}$$

where $\Pi_+(q)$ is the set of positive finite measures π_+ on $\mathcal{A}$ such that $\mathbb{E}_{\pi_+}[\Psi(\mu)-q]=0$ and $\mathbb{E}_{\pi_+}[\mathbb{D}(\mu)[B]]=1$. To that end, first observe that

$$\mathcal{U}(\Pi(q)|B)=\sup_{\pi\in\Pi(q):\pi\odot\mathbb{D}[B]>0}\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big]$$

and that, for any $\pi\in\Pi(q)$ such that $\pi\odot\mathbb{D}[B]>0$,

$$\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big]=\frac{\mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\mathbb{D}(\mu)[B]\big]}{\mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big]}.\tag{91}$$

Now consider $\pi\in\Pi(q)$ such that $\pi\odot\mathbb{D}[B]>0$. Then $\pi_+:=\pi/\mathbb{E}_\pi[\mathbb{D}(\mu)[B]]$ is an element of $\Pi_+(q)$ and (91) implies that

$$\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big]=\mathbb{E}_{\mu\sim\pi_+}\big[\Phi(\mu)\mathbb{D}(\mu)[B]\big].$$

Conversely, if $\pi_+\in\Pi_+(q)$, then $\pi:=\pi_+/\pi_+[\mathcal{A}]$ is an element of $\Pi(q)$ such that $\pi\odot\mathbb{D}[B]>0$ and

$$\mathbb{E}_{\mu\sim\pi_+}\big[\Phi(\mu)\mathbb{D}(\mu)[B]\big]=\frac{\mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\mathbb{D}(\mu)[B]\big]}{\mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big]}.$$

Since the above argument also shows that $\Pi(q)\odot_B\mathbb{D}$ is nonempty if and only if $\Pi_+(q)$ is nonempty, (90) follows. The right hand side of (90) is a linear program in π_+, so Theorem 4.7 implies that the supremum in π_+ can be achieved by assuming π_+ to be the weighted sum of at most $n+1$ Dirac masses, i.e. by assuming that

$$\pi_+=\sum_{i=0}^{n}\alpha_i\delta_{\mu_i}.\tag{92}$$

This finishes the proof of Theorem 4.8. □

8.7. Proof of Theorem 4.11

First let us show that, for $\lambda\in\mathbb{R}$, the statement that

$$\mathbb{E}_{\pi\odot\mathbb{D}}[\Phi|B]>\lambda,\quad \pi\in(\Psi^{-1}(\mathfrak{Q}))_B\tag{93}$$

is equivalent to the statement that

$$\mathbb{E}_{\mu\sim\pi}\Big[\big(\Phi(\mu)-\lambda\big)\mathbb{D}(\mu)[B]\Big] > 0. \tag{94}$$

To that end, assume (93) and observe that the definition (45) of $(\Psi^{-1}(\mathfrak{Q}))_B$ implies that $\pi\cdot\mathbb{D}[B]>0$, where, by (46),

$$\pi\cdot\mathbb{D}[B] := \mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big]. \tag{95}$$

Consequently, by (43),

$$\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big] = \frac{\mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\mathbb{D}(\mu)[B]\big]}{\mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big]} > \lambda,$$

and the denominator is strictly positive. Therefore,

$$\begin{aligned}&\mathbb{E}_{\mu\sim\pi}\big[\big(\Phi(\mu)-\lambda\big)\mathbb{D}(\mu)[B]\big]\\ &\quad= \mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\mathbb{D}(\mu)[B]\big] - \lambda\mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big]\\ &\quad> 0,\end{aligned}$$

and (94) follows. Conversely, assume (94) and observe that $\pi\cdot\mathbb{D}[B]>0$. To see this, observe that, if $\pi\cdot\mathbb{D}[B]=0$, then (95) implies that $\mathbb{D}(\mu)[B]=0$ π-a.s. and so

$$\mathbb{E}_{\mu\sim\pi}\big[\big(\Phi(\mu)-\lambda\big)\mathbb{D}(\mu)[B]\big] = 0,$$

which is a contradiction. Consequently, $\pi\cdot\mathbb{D}[B]>0$ and dividing the assumption

$$\begin{aligned}&\mathbb{E}_{\mu\sim\pi}\big[\big(\Phi(\mu)-\lambda\big)\mathbb{D}(\mu)[B]\big]\\ &\quad= \mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\mathbb{D}(\mu)[B]\big] - \lambda\mathbb{E}_{\mu\sim\pi}\big[\mathbb{D}(\mu)[B]\big]\\ &\quad> 0\end{aligned}$$

by $\pi\cdot\mathbb{D}[B] := \mathbb{E}_{\mu\sim\pi}[\mathbb{D}(\mu)[B]]$ throughout yields (93) and the equivalence is established.

The main assertion now follows from a direct application of Theorem 3.11. Finally, since Φ is semibounded, it follows that $\mu\mapsto\Phi(\mu)\mathbb{D}(\mu)[B]$ is semibounded and measurable, and the measure-affinity assertion follows from Lemma A.9. □

8.8. Proof of Theorem 4.13

Let us first establish that the assumptions of the theorem are well defined. To that end, note that Lemma 3.10 implies that $q\mapsto\inf_{\mu\in\Psi^{-1}(q)}\mathbb{D}(\mu)[B]$ is $\widehat{\mathcal{B}}(\mathrm{supp}(\mathbb{Q}))$-measurable and hence (58) is well defined. Similarly (59) is well defined.

For the proof of the theorem, fix $\delta>0$, let $\mathbb{Q}\in\mathfrak{Q}$ and $\mathbb{D}\in\mathfrak{D}$ satisfy the assumptions, and define $\lambda := \mathcal{U}(\mathcal{A})-\delta$. Since $(\Phi(\mu)-\lambda)\mathbb{D}(\mu)[B]$ is bounded and measurable, Lemma 3.10 implies that the function $q\mapsto\theta(q) :=$

$\sup_{\mu\in\Psi^{-1}(q)}(\Phi(\mu)-\lambda)\mathbb{D}(\mu)[B]$ is $\widehat{\mathcal{B}}(\mathrm{supp}(\mathbb{Q}))$-measurable. Moreover, (58) implies that the function θ is non-negative with $\mathbb{Q}$-probability one and (59) implies that θ is strictly positive on a subset of strictly positive $\mathbb{Q}$-measure. Hence,

$$E_{q\sim\mathbb{Q}}\left[\sup_{\mu\in\Psi^{-1}(q)}(\Phi(\mu)-\lambda)\mathbb{D}(\mu)[B]\right]=\mathbb{E}_{\mathbb{Q}}[\theta]>0,$$

and, therefore,

$$\sup_{Q\in\mathfrak{Q}}\mathbb{E}_{q\sim\mathbb{Q}}\left[\sup_{\mu\in\Psi^{-1}(q)}(\Phi(\mu)-\lambda)\mathbb{D}(\mu)[B]\right]>0.$$

It then follows from Theorem 4.11 that $\mathcal{U}(\Psi^{-1}\mathfrak{Q}|B)\geq\lambda=\mathcal{U}(\mathcal{A})-\delta$. Since $\delta>0$ was arbitrary, it follows that $\mathcal{U}(\Psi^{-1}\mathfrak{Q}|B)\geq\mathcal{U}(\mathcal{A})$. Theorem 4.5 implies that

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}|B)\leq\mathcal{U}(\mathcal{A})$$

and the theorem follows. □

8.9. Proof of Theorem 6.1

We will need the following notation: for an admissible set $\Pi\subseteq\mathcal{M}(\mathcal{A})$ of priors, an observation map $\mathbb{D}$, and an open subset $B\subseteq\mathcal{D}$, let $\Pi\odot_B\mathbb{D}$ be the set of probability distributions $\pi\odot\mathbb{D}$ on $\mathcal{A}\times\mathcal{D}$ generated by $\pi\in\Pi$:

$$\Pi\odot_B\mathbb{D}:=\{\pi\odot\mathbb{D}\mid\pi\in\Pi\text{ and }(\pi\cdot\mathbb{D})[B]>0\}.$$

We also define

$$\mathcal{U}(\Pi\odot_B\mathbb{D}):=\sup_{\pi\odot\mathbb{D}\in\Pi\odot_B\mathbb{D}}\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big].$$

The proof follows from the proof of Theorem 4.13 as follows. Let $\delta>0$, and let a measurable section ψ satisfy the assumptions. Define $\lambda:=\mathfrak{Q}^{\infty}(\Phi\circ\psi)-\delta$, and the universally measurable function $q\mapsto\theta(q):=\sup_{\mu\in\Psi^{-1}(q)}(\Phi(\mu)-\lambda)\mathbb{D}(\mu)[B]$. Then assumption (73) implies that the function θ is non-negative. It follows from the definition (86) of $\mathfrak{Q}^{\infty}(\Phi\circ\psi)$, and $\lambda<\mathfrak{Q}^{\infty}(\Phi\circ\psi)$, that there is a $\mathbb{Q}\in\mathfrak{Q}$ such that $\Phi\circ\psi>\lambda$ with nonzero $\mathbb{Q}$-measure. Since

$$\begin{aligned}\theta(q)&=\sup_{\mu\in\Psi^{-1}(q)}(\Phi(\mu)-\lambda)\mathbb{D}(\mu)[B]\\&\geq(\Phi\circ\psi(q)-\lambda)\mathbb{D}(\psi(q))[B],\end{aligned}$$

the positivity assumption (74) implies that θ is positive on a subset of positive $\mathbb{Q}$-measure. Hence,

$$\mathbb{E}_{q\sim\mathbb{Q}}\left[\sup_{\mu\in\Psi^{-1}(q)}(\Phi(\mu)-\lambda)\mathbb{D}(\mu)[B]\right]=\mathbb{E}_{\mathbb{Q}}[\theta]>0$$

and, therefore,

$$\sup_{\mathbb{Q}\in\mathfrak{Q}} \mathbb{E}_{q\sim\mathbb{Q}}\left[\sup_{\mu\in\Psi^{-1}(q)} (\Phi(\mu)-\lambda)\mathbb{D}(\mu)[B]\right] > 0.$$

It then follows from Theorem 4.11 that

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}|B) \geq \lambda = \mathfrak{Q}^{\infty}(\Phi\circ\psi) - \delta.$$

Since $\delta > 0$ was arbitrary, the assertion is proved. □

8.10. Proof of Theorem 6.9

We appeal to the corollary, Theorem 6.1, to Theorem 4.13. To that end, let $\mathcal{A}$ be defined as in (80), and let $\mathcal{Q} := \mathcal{A}_0$, $\Psi := P_0$, $\mathfrak{D} := \{\mathbb{D}^n\}$.

Since $\mathbb{D}^n = \mathbb{D}_0^n \circ P_\alpha$ is a pull-back,

$$\begin{aligned}(\pi\cdot\mathbb{D}^n)[B_\delta^n] &= \mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\mathbb{D}^n(\mu_1,\mu_2)[B_\delta^n]\big]\\ &= \mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\mathbb{D}_0^n\circ P_\alpha(\mu_1,\mu_2)[B_\delta^n]\big]\\ &= \mathbb{E}_{\mu_2\sim P_\alpha\pi}\big[\mathbb{D}_0^n(\mu_2)[B_\delta^n]\big]\\ &= (P_\alpha\pi\cdot\mathbb{D}_0^n)[B_\delta^n],\end{aligned}$$

from which we conclude that $(P_\alpha\pi\cdot\mathbb{D}_0^n)[B_\delta^n] > 0$ if and only if $(\pi\cdot\mathbb{D}^n)[B_\delta^n] > 0$, and so conclude

$$\Pi_\alpha \odot_{B_\delta^n} \mathbb{D}_0^n = P_\alpha\big(\Pi\odot_{B_\delta^n}\mathbb{D}^n\big),$$

where P_α acts on each component in the natural way. Moreover since $\Phi = \Phi_0\circ P_\alpha$ is also a pull-back, for $\pi\in\Pi$, we have

$$\begin{aligned}\mathbb{E}_{\pi\odot\mathbb{D}^n}\big[\Phi\big|B_\delta^n\big] &= \frac{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\Phi(\mu_1,\mu_2)\mathbb{D}^n(\mu_1,\mu_2)[B_\delta^n]\big]}{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\mathbb{D}^n(\mu_1,\mu_2)[B_\delta^n]\big]}\\ &= \frac{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\Phi_0\circ P_\alpha(\mu_1,\mu_2)\cdot\mathbb{D}_0^n\circ P_\alpha(\mu_1,\mu_2)[B_\delta^n]\big]}{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\mathbb{D}_0^n\circ P_\alpha(\mu_1,\mu_2)[B_\delta^n]\big]}\\ &= \frac{\mathbb{E}_{\mu_2\sim P_\alpha\pi}\big[\Phi_0(\mu_2)\mathbb{D}_0^n(\mu_2)[B_\delta^n]\big]}{\mathbb{E}_{\mu_2\sim P_\alpha\pi}\big[\mathbb{D}_0^n(\mu_2)[B_\delta^n]\big]}\\ &= \mathbb{E}_{P_\alpha\pi\odot\mathbb{D}_0^n}\big[\Phi_0\big|B_\delta^n\big]\end{aligned}$$

and so we conclude that

$$\mathcal{U}(\Pi\odot_{B_\delta^n}\mathbb{D}^n) = \mathcal{U}(\Pi_\alpha\odot_{B_\delta^n}\mathbb{D}_0^n). \tag{96}$$

We will now need the following proposition

Proposition 8.1. *Define the total variation metric d_{TV} on $\mathcal{M}(\mathcal{X})$ by*

$$d_{\mathrm{TV}}(\mu_1,\mu_2) := \sup_{A\in\mathcal{B}(\mathcal{X})} |\mu_1(A)-\mu_2(A)|.$$

Consider $B \in \mathcal{B}(\mathcal{X})$. Then for $\mu \in \mathcal{M}(\mathcal{X})$ such that $\mu(B) < 1$, we have

$$d_{\mathrm{TV}}(\mu, \mu|_{B^c}) \leq \mu(B).$$

Proof. For $A \in \mathcal{B}(\mathcal{X})$, we have

$$\begin{aligned}\mu(A) - \mu|_{B^c}(A) &= \mu(A) - \frac{\mu(A \cap B^c)}{\mu(B^c)} \\ &= \mu(A \cap B) + \mu(A \cap B^c) - \frac{\mu(A \cap B^c)}{\mu(B^c)} \\ &= \mu(A \cap B) - \frac{\mu(B)}{1 - \mu(B)}\mu(A \cap B^c)\end{aligned}$$

and therefore

$$\mu(A) - \mu|_{B^c}(A) \leq \mu(A \cap B) \leq \mu(B)$$

and

$$\begin{aligned}\mu(A) - \mu|_{B^c}(A) &\geq -\frac{\mu(B)}{1 - \mu(B)}\mu(A \cap B^c) \\ &\geq -\frac{\mu(B)}{1 - \mu(B)}\mu(B^c) \\ &= -\mu(B),\end{aligned}$$

thus establishing the assertion. □

For $B \in \mathcal{B}(\mathcal{X})$ and $\mu \in \mathcal{M}(\mathcal{X})$ such that $\mu(B) < 1$, the conditional measure $\mu|_{B^c} \in \mathcal{M}(\mathcal{X})$ is defined by

$$\mu|_{B^c}(A) := \frac{\mu(A \cap B^c)}{\mu(B^c)}, \quad A \in \mathcal{B}(\mathcal{X}).$$

It follows from Proposition 8.1 that $d_{\mathrm{TV}}(\mu, \mu|_{B^c}) \leq \mu(B)$ and since $d_{\mathcal{M}} \leq d_{\mathrm{TV}}$ (see e.g. [60, Eq. 2.24]), we conclude that

$$d_{\mathcal{M}}(\mu, \mu|_{B^c}) \leq \mu(B). \tag{97}$$

Let $B_\delta := B_\delta(x_1)$ denote the ball about the first sample of $x^n = (x_1, \ldots, x_n)$. Then, for $\mu_0 \in \mathcal{A}_0$, it follows from the assumptions that

$$\begin{aligned}d_{\mathcal{M}}(\mu_0, \mu_0|_{B_\delta^c}) &\leq \mu_0(B_\delta) \\ &\leq \mathcal{P}^\infty(\delta) \\ &< \alpha\end{aligned}$$

and therefore

$$\left(\mu_0, \mu_0|_{B_\delta^c}\right) \in \Psi^{-1}\mu_0.$$

Moreover, since

$$\mathbb{D}^n_{(\mu_0, \mu_0|_{B_\delta^c})}[B_\delta^n] = \left(\mu_0|_{B_\delta^c}\right)^n[B_\delta^n]$$

$$\begin{aligned}&\leq \mu_0|_{B_\delta^c}[B_\delta]\\&= 0,\end{aligned}$$

we conclude that the condition (73)

$$\inf_{(\mu_0,\mu_0')\in\Psi^{-1}\mu_0} \mathbb{D}^n(\mu_0,\mu_0')[B_\delta^n] = 0$$

of Theorem 6.1 is satisfied for all $\mu_0 \in \mathcal{A}_0$.

Now consider the diagonal map $\Delta\colon \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ defined by

$$\Delta(\mu) := (\mu,\mu), \quad \mu \in \mathcal{M}(\mathcal{X}).$$

Since

$$\Psi \circ \Delta(\mu) = P_0 \circ \Delta(\mu) = \mu, \quad \text{for all } \mu \in \mathcal{M}(\mathcal{X}),$$

it follows, if we define Δ on the first component of the product $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ and then restrict to $\mathcal{A}_0$, that Δ is a section of $\Psi = P_0$. It is clearly measurable, but also satisfies

$$P_\alpha \circ \Delta(\mu) = \mu, \quad \text{for all } \mu \in \mathcal{M}(\mathcal{X}),$$

that is, $P_\alpha \circ \Delta$ is the identity map from the first component of $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ to the second. Then, for $\mu_0 \in \mathcal{A}_0$, the positivity of the model $\mathcal{P}$ implies that

$$\begin{aligned}\mathbb{D}^n\big(\Delta(\mu_0)\big)[B_\delta^n] &= \mathbb{D}_0^n \circ P_\alpha\big(\Delta(\mu_0)\big)[B_\delta^n]\\&= \mathbb{D}_0^n(\mu_0)[B_\delta^n]\\&= (\mu_0)^n[B_\delta^n]\\&= \prod_{i=1}^n \mu_0[B_\delta(x_i)]\\&> 0\end{aligned}$$

so that the second condition (74) of Theorem 6.1 is satisfied for all $\mu_0 \in \mathcal{A}_0$. Theorem 6.1 then asserts that

$$\mathcal{U}(\Psi^{-1}\Pi_0 \odot_{B_\delta^n} \mathbb{D}^n) \geq \Pi_0^\infty(\Phi \circ \Delta).$$

Moreover, since

$$\Phi \circ \Delta = \Phi_0 \circ P_\alpha \circ \Delta = \Phi_0,$$

now as a function on the first component of $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$, and

$$\Psi^{-1}\Pi_0 = P_0^{-1}\Pi_0 = \Pi,$$

we conclude that

$$\mathcal{U}(\Pi \odot_{B_\delta^n} \mathbb{D}^n) \geq \Pi_0^\infty(\Phi_0).$$

The identity $\mathcal{U}(\Pi \odot_{B_\delta^n} \mathbb{D}^n) = \mathcal{U}(\Pi_\alpha \odot_{B_\delta^n} \mathbb{D}_0^n)$ of (96) then implies the assertion. □

Appendix

The following lemma is Lemma III.39 p. 86 of [29]. We also refer to p. 87 of [29] for the existence of the measurable selection η (which is also derived from Theorem III.38 p.85 of [29]). These results are related to Aumann's measurable section principle [7] (the extension to Suslin space is due to Sainte-Beuve [85]).

Lemma A.1. *Let $(T, \mathcal{T})$ be a measurable space, S a Suslin space. $\varphi\colon T \times S \to \bar{R}$ a $\mathcal{T} \otimes \mathcal{B}(S)$ measurable function and Γ a multifunction (i.e. a set-valued map) from T to non-empty subsets of S whose graph G belongs to $\mathcal{T} \times \mathcal{B}(S)$. Then*

1. *the function*
$$m(t) := \sup\{\phi(t, x) \mid x \in \Gamma(t)\}$$
is a $\widehat{\mathcal{T}}$-measurable function of t.
2. *for $\delta > 0$, there exists η, a $\widehat{\mathcal{T}}$-measurable function of t, such that $\eta(t) \in \Gamma(t)$ and $\varphi(t, \eta(t)) > m(t) - \delta$.*

The following definition is Definition 4.50 in [4]:

Definition A.2. Let (S, Σ) be a measurable space, and let X and Y be topological spaces. A function $h\colon S \times X \to Y$ is a *Carathéodory function* if:

1. for each $x \in X$, the function $h^x = h(., x)\colon S \to Y$ is $(\Sigma, \mathcal{B}(Y))$-measurable; and
2. for each $s \in S$, the function $h_s = h(s, .)\colon X \to Y$ is continuous.

The following lemma is Lemma 4.51 in [4] (see also [29, p. 70]):

Lemma A.3. *Let (S, Σ) be a measurable space, X a separable metrizable space, and Y a metrizable space. Then every Carathéodory function $h\colon S \times X \to Y$ is jointly measurable.*

A.1. Universally measurable functions

For a topological space T let $\widehat{\mathcal{B}}(T)$ denote the σ-algebra of universally measurable sets. For a measure μ, let $\widehat{\mu}$ denote its completion. Here we state the following proposition that allows us to define the expected value of $\widehat{\mathcal{B}}(T)$ measurable functions with respect to Borel measures. In all statements in the following proposition, the assertions follow when the integrals involved exist, in particular for semibounded functions. The proof is straightforward but tedious and follows from e.g. [39, Thm. pg. 37], the English version of [34, Ch. 2, pg. 49], and [29].

Proposition A.4. *Let T be a topological space. Then we have*

- *For a measurable function f we have $\mathbb{E}_{\widehat{\mu}} f = \mathbb{E}_{\mu} f$*
- *Let f be $\widehat{\mathcal{B}}(T)$-measurable. Then there exist two measurable functions $\underline{f}$ and $\overline{f}$ such that*
$$\underline{f} \leq f \leq \overline{f}, \quad \mu(\underline{f} \neq \overline{f}) = 0$$

and, for any such functions, we have

$$\mathbb{E}_\mu[\underline{f}] = \mathbb{E}_{\widehat{\mu}}[f] = \mathbb{E}_\mu[\overline{f}]$$

- *For a fixed μ, $f \mapsto \mathbb{E}_{\widehat{\mu}}[f]$ defines an affine function on the cone of non-negative $\widehat{\mathcal{B}}(T)$-measurable functions*
- *For a fixed $\widehat{\mathcal{B}}(T)$-measurable function f, the function $\mathcal{M}(T) \ni \mu \mapsto \mathbb{E}_{\widehat{\mu}}[f]$ is affine.*
- *Suppose that f_1, f_2 are $\widehat{\mathcal{B}}(T)$-measurable non-negative functions such that $f_1 \leq f_2$. Then $\mathbb{E}_{\widehat{\mu}}[f_1] \leq E_{\widehat{\mu}}[f_2]$ for all $\mu \in \mathcal{M}(T)$.*

Proposition A.4 leads to the following definition for the expectation of $\widehat{\mathcal{B}}(T)$-measurable functions with respect to Borel probability measures on T:

Definition A.5. For a Borel probability measure $\mu \in \mathcal{M}(T)$, we define the integral of a $\widehat{\mathcal{B}}(T)$-measurable function f by

$$\mathbb{E}_\mu[f] := \mathbb{E}_{\widehat{\mu}}[f]$$

when the latter exists, where $\widehat{\mu}$ is the completion of the measure μ as described in [39, p. 37].

Recall that a *carrier* T for a probability measure $\mathbb{Q} \in \mathcal{M}(\mathcal{Q})$ is a set $T \in \mathcal{B}(\mathcal{Q})$ such that $\mathbb{Q}(T) = 1$. For a carrier T, since $T \in \mathcal{B}(\mathcal{Q})$, it follows that $\mathcal{B}(T) = \mathcal{B}(\mathcal{Q}) \cap T$ and we can define the trace measure $\mathbb{Q}_T \in \mathcal{M}(T)$ by $\mathbb{Q}_T(A) := \mathbb{Q}(A), A \in \mathcal{B}(\mathcal{Q}) \cap T$. The following proposition shows that the expectation of a function can be defined with respect to measures which possess carriers upon which the function is universally measurable:

Proposition A.6. *Let S be a topological space, and suppose that f is $\widehat{\mathcal{B}}(T)$-measurable for each measurable $T \subseteq S$. Suppose also that $\mathbb{Q} \in \mathcal{M}(S)$ has a carrier $T \subseteq S$. Then, using Definition A.5, any such carrier T defines an expectation*

$$\mathbb{E}_{\mathbb{Q}_T}[f] := \mathbb{E}_{\widehat{\mathbb{Q}}_T}[f],$$

and this definition is independent of the carrier; that is, if $T' \subset S$ is another carrier, then

$$\mathbb{E}_{\widehat{\mathbb{Q}}_{T'}}[f] = \mathbb{E}_{\widehat{\mathbb{Q}}_T}[f].$$

Moreover, this expectation satisfies the assertions of affinity and monotonicity of Proposition A.4.

We also need a change of variables formula for expectations of universally measurable functions.

Proposition A.7. *Let X and Y be topological spaces, $\Psi\colon X \to Y$ a measurable map and suppose that $f\colon Y \to \mathbb{R}$ is $\widehat{\mathcal{B}}(Y)$ measurable. Then $f \circ \Psi\colon X \to \mathbb{R}$ is $\widehat{\mathcal{B}}(X)$-measurable and, for $\pi \in \mathcal{M}(X)$,*

$$\mathbb{E}_{\Psi\pi}[f] = \mathbb{E}_\pi[f \circ \Psi].$$

For Suslin space $\mathcal{X}$ and a subset $M \subset \mathcal{M}(\mathcal{X})$ let $\Sigma(M)$ denote the smallest σ-subalgebra of subsets of M for which the the evaluation map $\nu \mapsto \nu(B)$ is measurable for all $B \in \mathcal{B}(\mathcal{X})$. The following version of a result of von Weizsacker & Winkler [97] as stated in [107, Thm. 3.1] will be useful to us:

Theorem A.8. *Let $\mathcal{X}$ be a Suslin space, let $f_1, \ldots, f_n \colon \mathcal{X} \to \mathbb{R}$ be measurable functions, and let $c_1, \ldots, c_n \in \mathbb{R}$ be given. Define*

$$H := \left\{ \nu \in \mathcal{M}(\mathcal{X}) \,\middle|\, f_i \text{ is } \nu\text{-integrable and } \mathbb{E}_\nu[f_i] \leq c_i, \text{ for } i = 1, \ldots, n \right\}$$

Then, for each $\nu \in H$, there is a probability measure p on $\Sigma(\mathrm{ext}(H))$ such that

$$\nu(B) = \int_{\mathrm{ext}(H)} \nu'(B) \, \mathrm{d}p(\nu'), \quad \text{for all } B \in \mathcal{B}(\mathcal{X}). \tag{98}$$

[107, Prop. 3.1] shows that if a measurable function $f \colon \mathcal{X} \to \mathbb{R}$ is integrable with respect to all measures in H (allowing the values ∞ and $-\infty$), then integration

$$F(\nu) := \int_{\mathcal{X}} f \, \mathrm{d}\nu$$

is measure affine per Definition 3.3. We need a slightly more general result:

Lemma A.9. *Consider the situation of Theorem A.8, let $f \colon \mathcal{X} \to \mathbb{R}$ be a semibounded universally measurable function. Then*

$$F(\nu) := \mathbb{E}_{\widehat{\nu}}[f], \quad \text{for } \nu \in H,$$

is measure affine per Definition 3.3.

The next lemma extends [107, Thm. 2.1] to the case where the constraint functions f_i, for $i = 1, \ldots, n$, are universally measurable:

Lemma A.10. *Let $\mathcal{X}$ be Suslin, and fix universally measurable real-valued functions $f_1, \ldots, f_n$ and constants $c_1, \ldots, c_n$. Then*

$$H := \left\{ \nu \in \mathcal{M}(\mathcal{X}) \,\middle|\, f_i \text{ is } \widehat{\nu}\text{-integrable and } \mathbb{E}_{\widehat{\nu}}[f_i] \leq c_i \text{ for } i = 1, \ldots, n \right\} \tag{99}$$

is convex and

$$\mathrm{ext}(H) = \left\{ \nu \in H \,\middle|\, \begin{array}{c} \nu = \sum_{i=1}^m \alpha_i \delta_{x_i}, \text{ where } m \leq n+1, \\ \alpha_i \geq 0, x_i \in \mathcal{X} \text{ for } i = 1, \ldots, m, \\ \sum_{i=1}^m \alpha_i = 1, 1 \leq m \leq n+1, \text{ and the vectors} \\ \big(f_1(x_i), f_2(x_i), \ldots, f_n(x_i), 1\big) \\ \text{for } 1 \leq i \leq m \text{ are linearly independent} \end{array} \right\}.$$

A.2. Proofs

A.2.1. Proof of Proposition A.6

Let T and T' be two carriers for $\mathbb{Q} \in \mathcal{M}(S)$ and f a function such that f_T and $f_{T'}$ are $\widehat{\mathcal{B}}(t)$- and $\widehat{\mathcal{B}}(T')$-measurable respectively. Then Proposition A.4 implies

that there are functions f_1, f_2 measurable on T and f_1', f_2' measurable on T' such that

$$\begin{aligned} f_1 \le f_T \le f_2 \qquad & \mathbb{Q}_T(f_1 \ne f_2) = 0 \\ f_1' \le f_{T'} \le f_2' \qquad & \mathbb{Q}_{T'}(f_1' \ne f_2') = 0 \end{aligned}$$

so that

$$\begin{aligned} \mathbb{E}_{\widehat{\mathbb{Q}}_T}[f_T] &= \mathbb{E}_{\mathbb{Q}_T}[f_1] \\ \mathbb{E}_{\widehat{\mathbb{Q}}_{T'}}[f_{T'}] &= \mathbb{E}_{\mathbb{Q}_{T'}}[f_1'] \end{aligned}$$

Now, it is easy to see that $T \cap T'$ is also a carrier and that we have

$$f_1(x) \le f(x) \le f_2(x), \quad x \in T \cap T'$$

and

$$\mathbb{Q}_{T\cap T'}(f_1 \ne f_2) \le \mathbb{Q}_T(f_1 \ne f_2) = 0$$

so that we conclude from Proposition A.4 that

$$\begin{aligned} \mathbb{E}_{\widehat{\mathbb{Q}}_{T\cap T'}}[f] &= \mathbb{E}_{\mathbb{Q}_{T\cap T'}}[f_1] \\ &= \mathbb{E}_{\mathbb{Q}_T}[f_1] - \mathbb{E}_{\mathbb{Q}_{T\setminus T'}}[f_1] \\ &= \mathbb{E}_{\mathbb{Q}_T}[f_1] \\ &= \mathbb{E}_{\widehat{\mathbb{Q}}_T}[f] \end{aligned}$$

and so conclude that

$$\mathbb{E}_{\widehat{\mathbb{Q}}_{T\cap T'}}[f] = \mathbb{E}_{\widehat{\mathbb{Q}}_T}[f].$$

By the same argument on T' we conclude that $\mathbb{E}_{\widehat{\mathbb{Q}}_{T\cap T'}}[f] = \mathbb{E}_{\widehat{\mathbb{Q}}_{T'}}[f]$ and therefore the first assertion is proved. The assertions of affinity and monotonicity are similarly straightforward. □

A.2.2. Proof of Proposition A.7

Consider $\pi \in \mathcal{M}(X)$ and its pushforward $\nu := \Psi\pi$. By Proposition A.4 and the assumptions, there exists two measurable functions $\underline{f}$ and $\overline{f}$ such that

$$\underline{f} \le f \le \overline{f}, \quad \nu(\underline{f} \ne \overline{f}) = 0$$

from which we conclude that

$$\underline{f} \circ \Psi \le f \circ \Psi \le \overline{f} \circ \Psi$$

and

$$\begin{aligned} 0 &= \nu[\underline{f} \ne \overline{f}] \\ &= \Psi\pi[\underline{f} \ne \overline{f}] \end{aligned}$$

$$\begin{aligned} &= \pi[\Psi^{-1}\{\underline{f} \neq \overline{f}\}] \\ &= \pi[\underline{f} \circ \Psi \neq \overline{f} \circ \Psi] \end{aligned}$$

so that we obtain

$$\pi[\underline{f} \circ \Psi \neq \overline{f} \circ \Psi] = 0.$$

Since π was arbitrary, it follows that $f \circ \Psi$ is $\widehat{\mathcal{B}}(X)$-measurable. To obtain the change of variables formula, compute

$$\begin{aligned} \mathbb{E}_\pi[f \circ \Psi] &:= \mathbb{E}_{\widehat{\pi}}[f \circ \Psi] \\ &= \mathbb{E}_\pi[\overline{f} \circ \Psi] \\ &= \mathbb{E}_{\Psi\pi}[\overline{f}] \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_{\Psi\pi}[f] &:= \mathbb{E}_{\widehat{\Psi\pi}}[f] \\ &= \mathbb{E}_{\Psi\pi}[\overline{f}] \end{aligned}$$

from which we conclude the change of variables formula

$$\mathbb{E}_\pi[f \circ \Psi] = \mathbb{E}_{\Psi\pi}[f],$$

which completes the proof. □

A.2.3. Proof of Lemma A.9

Fix $\nu \in H$ and a probability measure p such that the barycentric formula (98) holds. Proposition A.4 asserts that there are measurable functions $f_1 \leq f \leq f_2$ such that $\nu(f_1 \neq f_2) = 0$. Therefore, $f_2 - f \geq 0$, $\mathbb{E}_{\widehat{\nu}}(f_2 - f) = 0$, $f - f_1 \geq 0$, and $\mathbb{E}_{\widehat{\nu}}(f - f_1) = 0$. Moreover, it is easy to see then we can make both f_1 and f_2 semibounded. Therefore F is a well defined extended real valued function. Moreover, [107, Prop. 3.1] asserts that the function $\nu \mapsto \mathbb{E}_\nu[f_i]$ is measure affine for $i = 1, 2$, and so

$$\mathbb{E}_\nu[f_i] = \int_{\mathrm{ext}(H)} \mathbb{E}_{\nu'}[f_i]\, \mathrm{d}p(\nu'), \quad \text{for } i = 1, 2.$$

Consequently, since $\nu[f_1 \neq f_2] = 0$, it follows that $\mathbb{E}_\nu[f_2 - f_1] = 0$ so that

$$0 = \mathbb{E}_\nu[f_2 - f_1] = \int_{\mathrm{ext}(H)} \mathbb{E}_{\nu'}[f_2 - f_1]\, \mathrm{d}p(\nu'), \quad \text{for } i = 1, 2,$$

and since $f_2 - f_1 \geq 0$ it follows that

$$\nu'[f_2 \neq f_1] = 0, \quad p\text{-a.e.}$$

and therefore

$$\widehat{\nu'}[f \neq f_1] = 0, \quad p\text{-a.e.}$$

Therefore we conclude that

$$\begin{aligned}
F(\nu) &:= \mathbb{E}_{\widehat{\nu}}[f] \\
&= \mathbb{E}_{\widehat{\nu}}[f_1] + \mathbb{E}_{\widehat{\nu}}[f - f_1] \\
&= \mathbb{E}_{\widehat{\nu}}[f_1] \\
&= \mathbb{E}_{\nu}[f_1] \\
&= \int_{\mathrm{ext}(H)} \mathbb{E}_{\nu'}[f_1]\,\mathrm{d}p(\nu') \\
&= \int_{\mathrm{ext}(H)} \mathbb{E}_{\widehat{\nu'}}[f_1]\,\mathrm{d}p(\nu') \\
&= \int_{\mathrm{ext}(H)} \mathbb{E}_{\widehat{\nu'}}[f_1]\,\mathrm{d}p(\nu') + \int_{\mathrm{ext}(H)} \mathbb{E}_{\widehat{\nu'}}[f - f_1]\,\mathrm{d}p(\nu') \\
&= \int_{\mathrm{ext}(H)} \mathbb{E}_{\widehat{\nu'}}[f]\,\mathrm{d}p(\nu') \\
&= \int_{\mathrm{ext}(H)} F(\nu')\,\mathrm{d}p(\nu'),
\end{aligned}$$

and the assertion is proved. □

A.2.4. Proof of Lemma A.10

Let us first establish that

$$\widehat{\nu_1 + \nu_2} = \widehat{\nu}_1 + \widehat{\nu}_2, \quad \text{for all } \nu_1, \nu_2 \in \mathcal{M}(\mathcal{X}), \tag{100}$$

$$\widehat{\alpha\nu} = \alpha\widehat{\nu}, \quad \text{for all } \nu \in \mathcal{M}(\mathcal{X}). \tag{101}$$

This follows from the fact that $(\nu_1 + \nu_2)(N) = 0$ if and only if $\nu_j(N) = 0$ for $j = 1, 2$ and the characterization of the completion $\widehat{\nu}$ by

$$\widehat{\nu}(B \cup S) := \nu(B), \quad B \in \mathcal{B}(\mathcal{X}),\ S \subset N,\ \nu(N) = 0$$

as found, for example, in [6, p. 18]. For then, for such B and S,

$$\begin{aligned}
\widehat{\nu_1 + \nu_2}(B \cup S) &= (\nu_1 + \nu_2)(B) \\
&= \nu_1(B) + \nu_2(B) \\
&= \widehat{\nu}_1(B \cup S) + \widehat{\nu}_2(B \cup S)
\end{aligned}$$

Now for the proof of the main assertion. Following the proof of [107, Thm. 2.1], it is sufficient to show that for

$$K := \left\{\nu \in \mathcal{M}(\mathcal{X}) \,\middle|\, f_i \text{ is } \widehat{\nu}\text{-integrable for } i = 1, \ldots, n\right\},$$

we have

$$\mathrm{ext}(K) := \{\delta_x, x \in \mathcal{X}\}, \tag{102}$$

and that $\mathbb{R}_+ K \subset \mathbb{R}_+ \mathcal{M}(\mathcal{X})$ is a lattice cone in its own ordering. For the first, observe that since $\mathrm{ext}(\mathcal{M}(\mathcal{X})) = \{\delta_x \mid x \in \mathcal{X}\}$ and that f_i are δ_x-integrable for

all $i = 1, \ldots, n$, $x \in \mathcal{X}$, it follows that

$$\{\delta_x \mid x \in \mathcal{X}\} \subseteq \operatorname{ext}(K).$$

Now suppose that $\nu \in \operatorname{ext}(K)$ is not a Dirac mass. Then, as in the proof that the extreme points of $\mathcal{M}(\mathcal{X})$ are the Dirac masses, see e.g. [4, Thm. 15.9], and using the fact that the support of ν must contain 2 or more points, we can decompose ν as a convex combination $\nu = \alpha\nu_1 + (1-\alpha)\nu_2$ where $\nu_1 \neq \nu_2$ and $\alpha \in (0,1)$. Moreover, from

$$\widehat{\nu} = \alpha\widehat{\nu_1} + (1-\alpha)\widehat{\nu_2},$$

we conclude that f_i being $\widehat{\nu}$-integrable implies that f_i is $\widehat{\nu_j}$-integrable for $j = 1, 2$ and $i = 1, \ldots, n$. Consequently, $\nu_j \in K$ for $j = 1, 2$. Since ν was an extreme point we conclude that $\nu_1 = \nu_2$ which is a contradiction, and (102) follows.

Now let us demonstrate that $\mathbb{R}_+K$ is a lattice cone in its own ordering. To that end, note that by [84, Lem. 10.4], it suffices to show that $\mathbb{R}_+K \subset \mathbb{R}_+\mathcal{M}(\mathcal{X})$ is a hereditary subcone, in that $\nu_1 \in \mathbb{R}_+K$, $\nu_2 \in \mathbb{R}_+\mathcal{M}(\mathcal{X})$ and $\nu_1 - \nu_2 \in \mathbb{R}_+K$ together imply that $\nu_2 \in \mathbb{R}_+K$. To that end, consider such ν_1 and ν_2. Then (100) implies that $\widehat{(\nu_1 - \nu_2)} = \widehat{\nu_1} - \widehat{\nu_2}$ and so we conclude that

$$0 \leq \mathbb{E}_{\widehat{(\nu_1-\nu_2)}}[|f_i|] = \mathbb{E}_{\widehat{\nu_1}}[|f_i|] - \mathbb{E}_{\widehat{\nu_2}}[|f_i|]$$

and therefore

$$\mathbb{E}_{\widehat{\nu_2}}[|f_i|] \leq \mathbb{E}_{\widehat{\nu_1}}[|f_i|] < \infty,$$

from which we conclude that $\nu_2 \in \mathbb{R}_+K$. Hence, $\mathbb{R}_+K$ is a hereditary subcone, and the assertion then follows as in the proof of [107, Thm. 2.1]. □

Acknowledgments

The authors gratefully acknowledge support for this work from the Air Force Office of Scientific Research under Award FA9550-12-1-0389 (Scientific Computation of Optimal Statistical Estimators). We thank P. Diaconis, D. Mayo, P. Stark, and L. Wasserman for stimulating discussions and relevant references and pointers. We thank the anonymous referees for detailed comments and suggestions.

References

[1] Abraham, C. and Cadre, B. (2002). Asymptotic properties of posterior distributions derived from misspecified models. *C. R. Math. Acad. Sci. Paris* **335** 495–498. MR1937120

[2] Abraham, C. and Cadre, B. (2008). Concentration of posterior distributions with misspecified models. *Ann. I.S.U.P.* **52** 3–14. MR2473284 (2010c:62014)

[3] Akhiezer, N. I. (1965). *The Classical Moment Problem and Some Related Questions in Analysis*. Hafner Publishing Co., New York. Translated by N. Kemmer. MR0184042 (32 ##1518)

[4] Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*, third ed. Springer, Berlin. MR2378491 (2008m:46001)
[5] Arveson, W. (1976). *An Invitation to C*-Algebras*. Springer-Verlag, New York.
[6] Ash, R. B. (1972). *Real Analysis and Probability. Probability and Mathematical Statistics, No. 11*. Academic Press, New York. MR0435320 (55 ##8280)
[7] Aumann, R. J. (1967). Measurable utility and the measurable choice theorem. *La décision C.N.R.S.* 15–26.
[8] Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.* **27** 1115–1122. MR0084241 (18,834b)
[9] Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718 (2000k:62097)
[10] Bauer, H. (2001). *Measure and Integration Theory. de Gruyter Studies in Mathematics* **26**. Walter de Gruyter & Co., Berlin. Translated from the German by Robert B. Burckel. MR1897176 (2003a:28001)
[11] Bayarri, M. J. and Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statist. Sci.* **19** 58–80.
[12] Belot, G. (2013). Bayesian orgulity. *Philos. Sci.* **80** 483–503. MR3135105
[13] Belot, G. (2013). Failure of calibration is typical. *arXiv:1306.4943*.
[14] Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.* **1** 385–402. MR2221271
[15] Berger, J. O. (1984). The robust Bayesian viewpoint. In *Robustness of Bayesian Analyses. Stud. Bayesian Econometrics* **4** 63–144. North-Holland, Amsterdam. With comments and with a reply by the author. MR785367
[16] Berger, J. O. (1994). An overview of robust Bayesian analysis. *Test* **3** 5–124. With comments and a rejoinder by the author. MR1293110 (95j:62018)
[17] Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37** 51–58; *correction, ibid* **37** 745–746. MR0189176 (32 ##6603)
[18] Berk, R. H. (1970). Consistency a posteriori. *Ann. Math. Statist.* **41** 894–906. MR0266356 (42 ##1262)
[19] Bernšteĭn, S. N. (1964). *Sobranie sochinenii. Tom IV: Teoriya veroyatnostei. Matematicheskaya statistika. 1911–1946*. Izdat. "Nauka", Moscow. MR0169758 (30 ##2)
[20] Bertsimas, D. and Popescu, I. (2005). Optimal inequalities in probability theory: a convex optimization approach. *SIAM J. Optim.* **15** 780–804 (electronic). MR2142860 (2006c:60020)
[21] Blei, D. M., Jordan, M. I. and Ng, A. Y. (2003). Hierarchical Bayesian models for applications in information retrieval. In *Bayesian Statistics, 7*

(Tenerife, 2002) 25–43. Oxford Univ. Press, New York. MR2003165

[22] Bogachev, V. I. (2007). *Measure Theory. Vol. II.* Springer-Verlag, Berlin.

[23] Bogachev, V. I. (2007). *Measure Theory. Vol. I.* Springer-Verlag, Berlin.

[24] Boole, G. (1854). *An Investigation of the Laws of Thought on Which Are Founded the Mathematical Theories of Logic and Probabilities*. Walton and Maberly, London.

[25] Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40** 318–335. MR0058937 (15,453e)

[26] Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. John Wiley & Sons Inc., New York. MR861118 (87m:62208)

[27] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge. MR2061575 (2005d:90002)

[28] Breiman, L., Le Cam, L. and Schwartz, L. (1964). Consistent estimates and zero-one sets. *Ann. Math. Statist.* **35** 157–161. MR0161413 (28 ##4620)

[29] Castaing, C. and Valadier, M. (1977). *Convex Analysis and Measurable Multifunctions. Lecture Notes in Mathematics, Vol. 580*. Springer-Verlag, Berlin. MR0467310 (57 ##7169)

[30] Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028. MR3127856

[31] Clarke, B. (2004). Comparing Bayes model averaging and stacking when model approximation error cannot be ignored. *J. Mach. Learn. Res.* **4** 683–712. MR2072265

[32] Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525 (94j:62010)

[33] Daley, D. J. and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes. Vol. II*, second ed. *Probability and Its Applications (New York)*. Springer, New York. General theory and structure. MR2371524 (2009b:60150)

[34] Dellacherie, C. and Meyer, P. A. (1975). *Probabilités et Potentiel*. Hermann, Paris. Chapitres I à IV, Édition entièrement refondue, Publications de l'Institut de Mathématique de l'Université de Strasbourg, No. XV, Actualités Scientifiques et Industrielles, No. 1372. MR0488194 (58 ##7757)

[35] Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–67. With a discussion and a rejoinder by the authors. MR829555 (88e:62016a)

[36] Diaconis, P. W. and Freedman, D. (1998). Consistency of Bayes estimates for nonparametric regression: normal theory. *Bernoulli* **4** 411–444. MR1679791 (2000b:62076)

[37] Donoho, D. L. (1988). One-sided inference about functionals of a density. *Ann. Statist.* **16** 1390–1420. MR964930 (89i:62057)
[38] Doob, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications. Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13* 23–27. Centre National de la Recherche Scientifique, Paris. MR0033460 (11,444c)
[39] Doob, J. L. (1994). *Measure Theory. Graduate Texts in Mathematics* **143**. Springer-Verlag, New York. MR1253752 (95c:28001)
[40] Dudley, R. M. (2002). *Real Analysis and Probability. Cambridge Studies in Advanced Mathematics* **74**. Cambridge University Press, Cambridge. Revised reprint of the 1989 original. MR1932358 (2003h:60001)
[41] Edwards, A. W. F. (1992). *Likelihood*, expanded ed. Johns Hopkins University Press, Baltimore.
[42] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review* **70** 193.
[43] Efron, B. (2013). Bayes' Theorem in the 21st Century. *Science* **340** 1177–1178.
[44] England and of Appeal (Civil Division), W. C. (2013). Nulty & Ors v. Milton Keynes Borough Council. [2013] EWCA Civ 15, Case No. A1/2012/0459. http://www.bailii.org/ew/cases/EWCA/Civ/2013/15.html.
[45] Feldman, J. (1958). Equivalence and perpendicularity of Gaussian processes. *Pacific J. Math.* **8** 699–708. MR0102760 (21 ##1546)
[46] Forrester, P. J. and Warnaar, S. O. (2008). The importance of the Selberg integral. *Bull Amer. Math. Soc.* **45** 489–534.
[47] Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27** 1119–1140. MR1740119 (2001g:62005)
[48] Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34** 1386–1403. MR0158483 (28 ##1706)
[49] Freedman, D. A. (1965). On the asymptotic behavior of Bayes estimates in the discrete case. II. *Ann. Math. Statist.* **36** 454–456. MR0174146 (30 ##4353)
[50] Fushiki, T. (2005). Bootstrap prediction and Bayesian prediction under misspecified models. *Bernoulli* **11** 747–758. MR2158259 (2006a:62118)
[51] Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Anal.* **3** 445–449. MR2434394
[52] Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In *Bayesian Nonparametrics. Camb. Ser. Stat. Probab. Math.* 35–79. Cambridge Univ. Press, Cambridge. MR2730660 (2011g:62020)
[53] Grünwald, P. D. (2006). Bayesian inconsistency under misspecification. http://homepages.cwi.nl/~pdg/ftp/valenciapost.pdf.
[54] Gustafson, P. and Wasserman, L. (1995). Local sensitivity diagnostics for Bayesian inference. *Ann. Statist.* **23** 2153–2167. MR1389870 (97e:62040)

[55] Guyon, I., Saffari, A., Dror, G. and Cawley, G. (2010). Model selection: beyond the Bayesian/Frequentist divide. *J. Mach. Learn. Res.* **11** 61–87.
[56] Hájek, J. (1958). On a property of normal distribution of any stochastic process. *Czechoslovak Math. J.* **8(83)** 610–618. MR0104290 (21 ##3045)
[57] Hausman, J. A. and Taylor, W. E. (1981). A generalized specification test. *Econom. Lett.* **8** 239–245. MR655491 (84j:62099)
[58] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415 (28 ##4622)
[59] Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics* 221–233. Univ. California Press, Berkeley, Calif. MR0216620 (35 ##7449)
[60] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*, second ed. *Wiley Series in Probability and Statistics*. John Wiley & Sons Inc., Hoboken, NJ. MR2488795 (2010j:62004)
[61] Johnstone, I. M. (2010). High dimensional Bernstein–von Mises: simple examples. In *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown. Inst. Math. Stat. Collect.* **6** 87–98. Inst. Math. Statist., Beachwood, OH. MR2798513 (2012k:62028)
[62] Kallenberg, O. (1975). *Random Measures*. Akademie-Verlag, Berlin. Schriftenreihe des Zentralinstituts für Mathematik und Mechanik bei der Akademie der Wissenschaften der DDR, Heft 23. MR0431372 (55 ##4372)
[63] Kechris, A. S. (1995). *Classical Descriptive Set Theory. Graduate Texts in Mathematics*. Springer-Verlag, New York.
[64] Kendall, D. G. (1962). Simplexes and vector lattices. *J. London Math. Soc.* **37** 365–371. MR0138983 (25 ##2423)
[65] Keynes, J. M. (1921). *A Treatise on Probability*. Macmillan and Co., London.
[66] Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877. MR2283395 (2009c:62074)
[67] Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Stat.* **6** 354–381.
[68] Kuznetsov, V. P. (1991). *Intervalnye Statisticheskie Modeli [Interval Statistical Models]*. "Radio i Svyaz'", Moscow. MR1186405 (93j:62001)
[69] Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.* **1** 277–329. MR0054913 (14,998b)
[70] Le Cam, L. and Schwartz, L. (1960). A necessary and sufficient condition for the existence of consistent estimates. *The Annals of Mathematical Statistics* 140–150.
[71] Leahu, H. (2011). On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat.* **5** 373–404. MR2802048 (2012g:62164)

[72] Lindley, D. V. (1985). *Making Decisions*, second ed. John Wiley & Sons, Ltd., London. MR892099 (88f:90009)
[73] Malakoff, D. (1999). Bayes offers a 'new' way to make sense of numbers. *Science* **286** 1460–1464.
[74] Martin, R. and Hong, L. (2012). On convergence rates of Bayesian predictive densities and posterior distributions. *arXiv:1210.0103v1*.
[75] Mayo, D. G. (2012). How can we cultivate Senn's ability? *RMM* **3** 14–18.
[76] Mayo, D. G. (2012). Statistical Science and Philosophy of Science Part 2: Shallow versus Deep Explorations. *RMM* **3**.
[77] Mayo, D. G. and Spanos, A. (2004). Methodology in practice: statistical misspecification testing. *Philos. Sci.* **71** 1007–1025 (2005). MR2133711 (2006b:62015)
[78] McGrayne, S. B. (2012). *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press.
[79] Nickl, R. (2013). Statistical Theory. http://www.statslab.cam.ac.uk/~nickl/Site/_files/stat2013.pdf.
[80] Owhadi, H. and Scovel, C. (2013). Brittleness of Bayesian inference and new Selberg formulas. Preprint at arXiv:1304.7046.
[81] Owhadi, H. and Scovel, C. (2014). Qualitative Robustness in Bayesian Inference. Preprint at arXiv:1411.3984.
[82] Owhadi, H., Scovel, C., Sullivan, T. J., McKerns, M. and Ortiz, M. (2013). Optimal uncertainty quantification. *SIAM Rev.* **55** 271–345.
[83] Oxtoby, J. C. (1971). *Measure and Category. A Survey of the Analogies Between Topological and Measure Spaces. Graduate Texts in Mathematics, Vol. 2*. Springer-Verlag, New York. MR0393403 (52 ##14213)
[84] Phelps, R. R. (2001). *Lectures on Choquet's Theorem*, second ed. *Lecture Notes in Mathematics* **1757**. Springer-Verlag, Berlin. MR1835574 (2002k:46001)
[85] Sainte-Beuve, M. F. (1974). On the extension of von Neumann-Aumann's theorem. *J. Functional Analysis* **17** 112–129. MR0374364 (51 ##10564)
[86] Schwartz, L. (1965). On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **4** 10–26. MR0184378 (32 ##1851)
[87] Schwartz, L. (1974). *Radon Measures on Arbitrary Topological Spaces and Cylindrical Measures*. Oxford Univ. Press, Oxford.
[88] Senn, S. (2007). Trying to be precise about vagueness. *Statistics in Medecine* **26** 1417–1430.
[89] Senn, S. (2011). You may believe you are a Bayesian but you are probably wrong. *RMM* **2** 48–66.
[90] Smith, J. E. (1995). Generalized Chebychev inequalities: theory and applications in decision analysis. *Oper. Res.* **43** 807–825. MR1361341 (96h:62013)

[91] Spanier, E. H. (1966). *Algebraic Topology*. Springer-Verlag, New York.
[92] Stuart, A. M. (2010). Inverse problems: a Bayesian perspective. *Acta Numer.* **19** 451–559. MR2652785
[93] Telgársky, R. V. (1987). Topological games: on the 50th anniversary of the Banach–Mazur game. *Rocky Mountain J. Math.* **17** 227–276. MR892457 (88d:54046)
[94] Tibshirani, R. and Wasserman, L. A. (1988). Sensitive parameters. *The Canadian Journal of Statistics* **16** 185–192.
[95] Topsøe, F. (1970). *Topology and Measure. Lecture Notes in Mathematics, Vol. 133*. Springer-Verlag, Berlin. MR0422560 (54 ##10546)
[96] von Mises, R. (1964). *Mathematical Theory of Probability and Statistics*. Academic Press, New York. Edited and Complemented by Hilda Geiringer. MR0178486 (31 ##2743)
[97] von Weizsäcker, H. and Winkler, G. (1979/80). Integral representation in the set of solutions of a generalized moment problem. *Math. Ann.* **246** 23–32. MR554129 (80m:46012)
[98] Walker, S. (2004). New approaches to Bayesian consistency. *Ann. Statist.* **32** 2028–2043. MR2102501 (2006c:62049)
[99] Walker, S. and Hjort, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 811–821. MR1872068 (2002i:62100)
[100] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities. Monographs on Statistics and Applied Probability* **42**. Chapman and Hall Ltd., London. MR1145491 (93d:62008)
[101] Wasserman, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In *Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statist.* **133** 293–304. Springer, New York. MR1630088 (99c:62146)
[102] Wasserman, L., Lavine, M. and Wolpert, R. L. (1993). Linearization of Bayesian robustness problems. *J. Statist. Plann. Inference* **37** 307–316. MR1248514 (94k:62046)
[103] Wasserman, L. and Seidenfeld, T. (1994). The dilation phenomenon in robust Bayesian inference. *J. Statist. Plann. Inference* **40** 345–356.
[104] Wasserman, L. A. (1990). Prior envelopes based on belief functions. *Ann. Statist.* **18** 454–464. MR1041404 (91b:62008)
[105] Weichselberger, K. (2000). The theory of interval-probability as a unifying concept for uncertainty. *Internat. J. Approx. Reason.* **24** 149–170. Reasoning with imprecise probabilities (Ghent, 1999). MR1766281 (2002e:68129)
[106] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR640163 (83b:62216)
[107] Winkler, G. (1988). Extreme points of moment sets. *Math. Oper. Res.* **13** 581–587. MR971911 (89i:60037)
[108] Zsilinszky, L. (1998). Topological games and hyperspace topologies. *Set-Valued Anal.* **6** 187–207. MR1646498 (99g:54005)